Apprendre ce qui ne peut pas varier : la mémoire comme contrainte du monde

Pourquoi les architectures prédictives latentes déplacent le problème de l'apprentissage sans franchir le seuil de la mémoire biographique

Quatrième volet de la série Encodage, transduction et modèles du monde. Les architectures de type JEPA n'apprennent pas à représenter le monde mais les contraintes qui rendent certaines transformations prédictibles. Ce déplacement est important, mais il opère à un niveau où la notion de mémoire biographique n'est pas définie.

1. Introduction : ce que la trilogie laisse ouvert

Les trois précédents articles de cette série (article 1, article 2, article 3) ont défendu une thèse progressivement étoffée. Toute architecture cognitive, biologique ou artificielle, opère par médiation représentationnelle. Les systèmes vivants accèdent à leurs représentations par une boucle perception-action dont les modèles de langage sont structurellement dépourvus. La différence la plus profonde entre cognition humaine et architectures contemporaines tient moins au contenu des représentations qu'à la nature des relations qui les relient : co-occurrence statistique d'un côté, **arête biographique** de l'autre.

Pour la suite, il importe de stabiliser une définition fonctionnelle de cette dernière notion, indépendamment de toute charge phénoménologique. Une arête biographique désigne une relation entre représentations mnésiques qui satisfait simultanément quatre conditions opératoires :

1. ***Indexation*** sur l'histoire d'un agent continu,
2. ***Co-activation*** dans un même épisode,
3. ***Préservation*** d'un contexte d'occurrence,
4. ***Possibilité de rappel situé*** depuis plusieurs perspectives modales.

Cette définition est délibérément fonctionnelle. Elle peut être tenue sans engagement préalable sur la nature subjective de l'expérience, et c'est précisément cette translation qui sera défendue en §11 contre l'objection qu'elle abolirait la position phénoménologique antérieure.

La question que la trilogie laisse ouverte se formule alors ainsi : si le déficit des grands modèles de langage tient à leur ancrage symbolique secondaire, et si l'incarnation robotique ne suffit pas à elle seule à produire une mémoire ainsi définie, qu'est-ce que les architectures prédictives auto-supervisées, JEPA et les architectures apparentées de prédiction latente, apportent réellement à ce paysage ?

La thèse défendue ici est la suivante : ces architectures déplacent l'objectif de l'apprentissage de manière épistémiquement décisive, en passant de la prédiction d'observations à la prédiction de contraintes. Ce déplacement est important. Il opère cependant à un niveau où les conditions de l'arête biographique ne sont pas définies. Non pas qu'il échoue à les satisfaire ; il ne les adresse pas.

Le titre assume une compression rhétorique : il ne s'agit pas d'apprendre des invariants absolus, mais ce qui, dans une distribution donnée, doit rester stable pour préserver la prédictibilité.

Le domaine de validité de cet article est restreint : on parle des architectures prédictives latentes auto-supervisées, dont JEPA est l'exemplaire le plus discuté, et de leurs extensions agentiques avec mémoire externe lorsque celles-ci se présentent comme dépassement du seuil identifié. Les conclusions ne

s'étendent pas mécaniquement aux LLM seuls, aux modèles génératifs purs, ni aux systèmes agentiques composites complets.

2. Trois niveaux à ne pas confondre

L'analyse qui suit opère sur trois niveaux qu'il importe de distinguer explicitement, faute de quoi les transitions deviennent ambiguës.

1. Le niveau ***computationnel*** concerne la mécanique des architectures : encodeurs, espaces latents, fonctions de prédiction, objectifs d'apprentissage, mécanismes de régularisation.
2. Le niveau ***épistémique*** concerne ce qui est appris au sens fort : structure de cohérence, invariants, contraintes capturées dans la représentation.
3. Le niveau ***phénoménologique*** concerne la mémoire vécue par un sujet situé, avec ses propriétés d'autonoéticité, de modulation affective et de navigabilité multi-perspective.

Ces trois niveaux ne sont pas substituables.

- Une propriété au niveau computationnel n'implique pas mécaniquement une propriété au niveau épistémique ;
- Une propriété épistémique n'implique pas mécaniquement une propriété phénoménologique.

Confondre les niveaux produit deux symétriques erreurs :

- Sur-attribuer (« JEPA comprend le monde »),
- Et sous-attribuer (« JEPA n'a pas de mémoire »).

JEPA n'échoue pas à produire une mémoire biographique. Il opère à un niveau où cette notion n'est pas définie.

Cette précision n'est pas une concession diplomatique. Elle conditionne ce qu'on est en droit d'attendre de ces architectures en environnement régulé, singulièrement en santé, où la robustesse hors distribution n'est pas un objectif secondaire mais une exigence de conformité.

3. L'erreur de cadrage : apprendre plus n'est pas apprendre mieux

Une lecture empiriste naïve, parfois associée, abusivement mais commodément, à une induction aristotélicienne, repose sur l'idée qu'à partir d'une accumulation suffisante d'observations, l'esprit dégage par abstraction les régularités du monde. Ce schéma a structuré la majeure partie du machine learning contemporain : davantage de données, davantage de paramètres, davantage de capacité, et la généralisation suivrait.

Les limites de cette posture sont aujourd'hui documentées avec une régularité presque embarrassante. Surapprentissage sur les corrélations superficielles, dépendance massive aux annotations, fragilité hors distribution, sensibilité aux ***shortcut features***. Ces phénomènes ne sont pas des bugs résiduels qui s'évanouiront au prochain doublement d'échelle. Ils sont la conséquence directe d'un objectif d'apprentissage mal posé : apprendre à reconstruire ce qui a été observé n'est pas apprendre ce qui structure les observations.

La thèse implicite des architectures prédictives latentes est plus précise : ***le problème n'est pas d'apprendre plus de données, mais de changer ce qui est appris***.

4. Le déplacement JEPA : généalogie et principe

JEPA n'arrive pas dans un paysage vide. Il s'inscrit dans une lignée d'architectures auto-supervisées qui a structuré la recherche en représentation depuis 2020 : les approches contrastives SimCLR [29] et MoCo [30], puis BYOL [31], qui montre qu'une prédiction asymétrique entre vues peut éviter l'apprentissage contrastif sans effondrement représentationnel, DINO [32] qui généralise la distillation, MAE [33] qui rétablit la reconstruction d'observations masquées comme objectif compétitif, et enfin I-JEPA [27] et V-JEPA [28] qui formalisent la prédiction latente sur paires contexte-cible.

La spécificité de JEPA par rapport à BYOL n'est pas l'idée d'une cible prédite dans l'espace latent (BYOL la portait déjà), mais le fait que la cible soit ***spatialement située*** via un signal de position, et que le contexte soit explicitement masqué plutôt que défini par augmentation. Ce détail change la nature de ce qui est appris : non plus une invariance à des transformations imposées, mais une prédictibilité conditionnée à une localisation.

Le principe formalisé par Yann LeCun dans son programme de 2022 [16] sur l'autonomous machine intelligence peut être énoncé brièvement. Un encodeur produit une représentation d'un contexte. Un second encodeur produit une représentation d'une cible partielle. Un prédicteur, opérant entièrement dans l'espace latent et conditionné sur la position de la cible, prédit la représentation de la cible à partir de celle du contexte. L'objectif d'apprentissage est défini sur la similarité entre prédiction et cible ***dans cet espace latent***, et non sur la reconstruction pixel-à-pixel de l'observation masquée.

Ce détail est conceptuellement décisif. Les autoencodeurs et les modèles de diffusion apprennent à reconstruire des observations brutes : ils sont contraints de représenter tout ce qui est dans le signal, y compris le bruit, les détails non pertinents, les variations de surface qui n'ont aucune valeur structurelle. Un JEPA ne cherche pas à reconstruire l'observation brute. Il prédit dans un espace de représentation appris, où les détails informationnellement inutiles ont été, au moins en principe, éliminés.

Une précision technique cruciale doit être faite ici, sous peine de transformer l'analyse en éloge sans fondement :

- Un objectif de prédictibilité latente seul converge vers la solution dégénérée où toutes les représentations s'effondrent en un point unique, le ***collapse***.
- Les propriétés d'invariance et de compression évoquées plus loin ne sont garanties que par des mécanismes anti-collapse explicites : régularisation par variance et covariance (VICReg [34]), target encoder mis à jour par moyenne mobile exponentielle (EMA), asymétrie d'architecture entre encodeur de contexte et encodeur de cible, ou combinaisons de ces dispositifs.
- Ce que JEPA apprend est défini par l'objectif ***combiné*** à ces contraintes structurelles, non par l'objectif seul.

Les vertus prédictives de l'architecture sont donc inséparables de choix d'ingénierie inductive qui doivent être documentés comme tels.

Le modèle n'apprend pas à voir. Il apprend ce qui est prévisible.

Cette formulation traduit un choix architectural précis : l'objectif d'apprentissage devient la ***cohérence inter-représentationnelle***, et non la fidélité observationnelle. Ce n'est pas un raffinement, c'est un changement de cible.

5. Ce que JEPA apprend réellement : contraintes de prédictibilité

Cette modification de cible a une conséquence souvent insuffisamment explicitée.

L'espace latent appris par un JEPA n'est pas un espace de ***features*** au sens classique du terme, c'est-à-dire un dictionnaire de motifs visuels ou sémantiques utiles pour des tâches en aval. C'est un espace de ***cohérence*** : une géométrie dans laquelle certaines configurations de représentations sont compatibles entre elles et d'autres ne le sont pas.

Une précision technique s'impose ici, car elle conditionne tout ce qui suit. JEPA standard, dans sa formulation I-JEPA et V-JEPA, n'encode pas explicitement les transformations du monde. Il n'est pas un modèle dynamique au sens d'un système qui simulerait des trajectoires d'états. Ce que l'apprentissage par prédiction latente sur des paires contexte-cible produit, c'est une fonction de mapping entre représentations latentes, contrainte de telle sorte que les paires correspondant à des co-occurrences naturelles dans les données aient des représentations mutuellement prédictibles. La dynamique du monde n'est pas représentée comme telle ; elle est implicitement ***contrainte par la structure de prédictibilité*** de l'espace latent.

JEPA n'encode pas les transformations elles-mêmes. Il encode les contraintes qui rendent certaines transformations prédictibles.

La nuance n'est pas cosmétique. Elle évite l'attribution facile d'une propriété dynamique explicite à un système qui reste, dans sa version standard, formellement statique : une fonction de mapping latent vers latent. Et elle situe correctement le rapport de JEPA aux ***world models*** explicites comme DreamerV3 [19] ou aux modèles d'inférence active de Friston [20] : ces derniers représentent explicitement des dynamiques itératives ; JEPA, lui, contraint un espace dans lequel certaines dynamiques deviennent prédictibles sans être simulées.

Cette caractérisation s'applique au JEPA standard. La roadmap originale de LeCun propose des extensions explicitement dynamiques, hiérarchiques (H-JEPA) ou conditionnées par l'action (A-JEPA), qui déplaceraient une partie de l'analyse présentée ici. Ces variantes restent à ce jour largement à l'état de programme, avec quelques implémentations partielles. Elles ne sont pas le sujet du présent article, mais leur existence interdit de figer la caractérisation « statique » de JEPA comme propriété définitionnelle de la famille architecturale entière.

La position que ces précisions permettent de tenir peut être formulée d'une seule phrase, qui sert de centre gravitationnel à l'ensemble de l'article. ***JEPA n'est ni une mémoire, ni un simulateur, ni un agent : c'est une architecture qui apprend une géométrie de prédictibilité***.

6. L'analogie mémoire longue : critique, et reconstruction positive

Une analogie circule régulièrement dans la littérature et dans les présentations de ces architectures. Le latent space d'un JEPA fonctionnerait comme une forme de ***mémoire longue***, voire comme un analogue approximatif de la proprioception, en ce qu'il maintient une cohérence interne stable à travers les transformations du signal d'entrée.

L'analogie capture une intuition partielle. Il y a effectivement, dans un JEPA bien entraîné, une forme de continuité représentationnelle : les transformations du signal d'entrée qui ne modifient pas la structure sous-jacente (occlusion partielle, bruit, transformations géométriques mineures) ne perturbent pas la représentation. Cette stabilité interne, indépendante des fluctuations de surface du signal, présente une parenté structurelle évidente avec les invariances perceptives décrites par la psychologie cognitive.

Mais l'analogie biologique rate l'essentiel. La proprioception biologique n'est pas une simple stabilité de représentation ; elle est l'émission continue, par le corps, d'un signal interne qui informe le système

nerveux central de l'état effectif du système moteur. Elle est ***incarnée*** au sens fort : il y a un corps, des récepteurs réels, une boucle sensorimotrice fermée. Aucun JEPA actuel ne possède quoi que ce soit de tel. Sa ***« cohérence »*** est purement représentationnelle ; elle n'est ancrée dans aucun substrat physiologique ni dans aucune action motrice.

La critique de l'analogie ne suffit cependant pas. Reste à dire ce que JEPA fait ***positivement***, sans béquille biologique. Trois propriétés méritent d'être nommées en propre.

1. D'abord, une ***invariance structurelle sous transformation partielle***. Les représentations apprises sont stables sous une classe de perturbations du signal d'entrée correspondant à des transformations qui préservent la structure prédictible des données. Cette invariance n'est pas posée a priori ; elle émerge de l'objectif d'apprentissage, modulo les contraintes anti-collapse rappelées en §4.

2. Ensuite, une ***compression orientée prédictibilité***. L'espace latent privilégie l'information qui contribue à la prédiction inter-représentationnelle, et marginalise l'information qui ne contribue pas. C'est une forme de filtrage informationnel par utilité prédictive, distinct du filtrage par perte d'information dans les autoencodeurs classiques.

3. Enfin, une ***sélection de l'information contrainte par la prédictibilité***. Les dimensions du signal d'entrée qui n'apportent aucune contrainte aux paires contexte-cible tendent à ne pas être préservées dans la représentation, non par décision attentionnelle d'un agent, mais par construction de l'objectif d'apprentissage. Ce qui n'est pas contraint par la prédictibilité n'a aucune raison d'être stabilisé dans la représentation.

Ces trois propriétés constituent un profil épistémique propre, qui n'a pas besoin de l'analogie biologique pour être caractérisé.

7. De l'observation à la continuation latente contrainte

Le déplacement opéré par JEPA peut alors être formulé en termes paradigmatiques. Le paradigme classique de l'apprentissage supervisé articule trois opérations : observer, abstraire, classer. Un échantillon entre, une catégorie sort. Le paradigme JEPA articule différemment : observer, contraindre, anticiper. Un contexte entre, un espace de cibles plausibles est défini.

La conséquence est précise : le modèle ***contraint un espace de continuations latentes plausibles, sans les expliciter***. Il ne génère pas de trajectoires complètes, il ne produit pas d'images photoréalistes, il ne déroule pas de simulations itératives. Mais sa structure d'espace latent rend certaines transitions prédictibles et d'autres non, ce qui revient à délimiter implicitement un espace de continuations admissibles. La distinction avec un simulateur explicite est cruciale : un ***world model*** dynamique génère des trajectoires ; JEPA délimite l'espace dans lequel ces trajectoires devraient pouvoir être générées par un système approprié.

Cette caractérisation rejoint, par un chemin entièrement différent, la littérature en ***predictive processing*** (Friston, Clark) : un cerveau prédictif n'est pas un cerveau qui génère des hallucinations, c'est un cerveau qui anticipe l'erreur de prédiction et minimise sa surprise libre. L'analogie a ses limites, l'inférence active suppose une boucle sensorimotrice qu'un JEPA standard n'a pas, mais elle situe correctement le type d'objet computationnel qu'on construit. Pas un classifieur, pas un générateur, pas un agent. Un système qui contraint un espace.

8. Les labels : une projection arbitraire dans un espace optimisé sur d'autres critères

Une conséquence pratique de ce déplacement concerne le statut des annotations supervisées. Le paradigme dominant les présente comme une ***vérité terrain***. Cette présentation contient une équivoque qu'il est utile de dissiper.

Les labels ne décrivent pas le monde. Ils contraignent son usage.

Le label « tumeur maligne » apposé sur une image radiologique ne décrit pas une propriété intrinsèque de l'image. Il indique l'usage clinique attendu de cette image dans un cadre de décision donné. La maladie n'est pas dans le pixel ; elle est dans l'articulation entre le pixel, l'histoire du patient, le protocole diagnostique et la décision thérapeutique. Le label compresse cette articulation en un signal binaire utile à l'apprentissage supervisé, mais cette compression est une projection métier, pas une description ontologique.

La tension avec JEPA est alors explicite et trop rarement formulée. Un JEPA apprend une géométrie de cohérence ***indépendamment de tout découpage métier***. L'espace latent qu'il construit est optimisé sur un critère interne, la prédictibilité inter-représentationnelle, ***qui n'a aucune raison structurelle de s'aligner sur les frontières taxonomiques d'un usage clinique particulier***. Lorsqu'on ajoute une tête supervisée à un encodeur SSL préentraîné, on ne ***complète*** pas l'apprentissage : on ***réinjecte*** dans un espace optimisé sur d'autres critères une projection arbitraire dictée par les besoins d'une tâche aval.

Cette réinjection est légitime. ***Elle est même indispensable pour les usages opérationnels***. Mais elle n'est pas neutre, et elle ne doit pas être confondue avec une révélation des structures que l'encodeur aurait découvertes : l'encodeur a découvert ***ses*** structures, et la tête supervisée projette ces structures sur les frontières d'une tâche donnée. D'où une partition stratégique : l'auto-supervision est un mécanisme de ***découverte de structure*** ; la supervision est un mécanisme de ***projection d'usage***. Les deux sont nécessaires. Ils ne font pas la même chose.

9. Implications en environnement régulé : santé

Cette distinction n'est pas spéculative. Elle a des conséquences techniques concrètes pour les systèmes IA en environnement régulé, et singulièrement en santé. Trois propriétés y prennent une importance particulière, qu'il faut formuler comme des ***hypothèses d'ingénierie*** et non comme des propriétés intrinsèques de l'architecture, mais qui disposent désormais d'un faisceau d'évidences empiriques substantiel.

1. Première hypothèse : la robustesse hors distribution. Un classifieur supervisé entraîné sur une cohorte hospitalière européenne, déployé sur une cohorte nord-américaine, voit ses performances se dégrader d'autant plus que ses représentations ont été optimisées pour des corrélations spécifiques au site d'entraînement. L'hypothèse opérationnelle est qu'un encodeur SSL préentraîné sur une distribution large, projeté ensuite par fine-tuning supervisé, se dégrade généralement moins. Cette hypothèse a reçu un soutien empirique en imagerie médicale notamment via les travaux d'Azizi et collaborateurs sur le SSL pour la classification d'images médicales [38], les architectures de type CheXzero [41] sur la radiographie thoracique, et plus récemment RETFound [39] pour l'imagerie ophtalmologique. Elle reste cependant à établir cas par cas : aucune garantie théorique ne l'impose, et l'écart de performance dépend fortement de la distance distributionnelle entre site d'entraînement et site de déploiement.

2. Deuxième hypothèse : la dépendance aux datasets annotés. La situation HDLSS (***High Dimension, Low Sample Size***), endémique en biomédecine, pénalise sévèrement la supervision pure. Le préentraînement auto-supervisé sur des corpus non annotés, lorsqu'il est

techniquement faisable, ***peut*** déplacer une partie de la complexité d'apprentissage hors de la phase coûteuse en annotation médicale. Cette possibilité dépend de la disponibilité d'un corpus de préentraînement structurellement comparable au domaine d'application. La revue de Krishnan et collaborateurs [40] sur le SSL en santé documente à la fois la promesse et les conditions restrictives de cette approche : pour des cohortes spécialisées rares (par exemple oncologie de précision sur sous-populations moléculaires), le corpus de préentraînement n'existe souvent pas dans le domaine cible, et l'usage de transferts depuis RadImageNet [42] ou autres référentiels génériques réintroduit des biais de domaine que la supervision finale doit traiter.

3. Troisième hypothèse : la modélisation de trajectoires. La médecine est essentiellement temporelle. Une architecture qui apprend des contraintes de prédictibilité sur des paires temporelles ***peut*** produire des représentations utiles pour modéliser des trajectoires de maladie ou de traitement. Là encore, il s'agit d'une hypothèse d'ingénierie, dont la vérification empirique pour les cohortes spécialisées reste à compléter.

Aucune des trois ne garantit son propre succès. Toutes trois orientent une stratégie d'ingénierie raisonnable lorsque les conditions sont remplies.

> ***Encadré illustratif***. *Dans le programme TweenMe / OCTOPUS sur mNSCLC porteurs de la mutation BRAF V600E (n=184, 5 pays européens), le travail sur les trajectoires a conduit à mobiliser une combinaison d'apprentissage de représentations et de modélisation par SurvTRACE [43], architecture transformer pour l'analyse de survie en présence d'événements compétitifs, avec une fidélité TSTR mesurée à 95,2 % sur la cohorte de validation, évaluée contre une baseline classifieur entraîné sur données réelles. Cette métrique n'est pas une preuve d'indistinguabilité statistique générale, ni une démonstration de la supériorité intrinsèque des représentations apprises. Elle est un indice, dans le cadre d'évaluation considéré, que la trajectoire générée préserve les propriétés opérationnelles utiles aux tâches aval. Terrain d'implémentation, pas démonstration universelle. Le détail méthodologique fait l'objet d'une publication séparée.*

10. Limites : ne pas transformer JEPA en religion

Plusieurs limites authentiques doivent être tenues fermement, au risque sinon de glisser dans le registre évangéliste qui guette toute architecture nouvelle.

1. D'abord, il n'existe à ce jour aucune preuve qu'un JEPA apprenne une *physique complète* du monde. Les démonstrations existantes (I-JEPA sur images [27], V-JEPA sur vidéos [28]) montrent des invariances apprises sur des distributions naturelles, mais ces invariances ne couvrent qu'une fraction des contraintes physiques réelles, et la généralisation à des régimes très éloignés de la distribution d'entraînement reste à démontrer.

2. Ensuite, le latent space appris est, dans la grande majorité des configurations, ***non interprétable***. Cette limite n'est pas spécifique à JEPA, elle est partagée par l'ensemble du SSL. Mais elle pèse particulièrement en contexte régulé où la traçabilité des features est exigée : un dispositif médical logiciel relevant du règlement MDR (UE) 2017/745 et, le cas échéant, du régime des systèmes d'IA à haut risque au titre du règlement européen sur l'IA, doit satisfaire des exigences cumulées de transparence, de robustesse et de supervision humaine. L'absence d'interprétabilité de l'espace latent doit alors être compensée par d'autres garanties : explicabilité post-hoc, monitoring de dérive, validation indépendante par cohorte externe, documentation technique étendue.

3. Troisièmement, l'évaluation de ces architectures est techniquement délicate. Les métriques classiques (précision, AUC) ne mesurent pas directement ce que JEPA est censé apprendre. Les ***probes*** linéaires sur tâches aval donnent une indication, pas une mesure directe de la qualité du monde modélisé.

4. Quatrièmement, la performance dépend fortement du ***design des masques*** et des ***stratégies de génération de paires contexte-cible***. Ce qui ressemble parfois à une découverte automatique est en réalité partiellement encodé dans des choix d'ingénierie inductive. Choix légitimes, mais qui doivent être documentés comme tels.

5. Cinquièmement, le préentraînement SSL nécessite un compute substantiel. L'argument de réduction de la dépendance aux annotations en §9 doit être lu en regard du fait que cette réduction se paye en cycles GPU sur des corpus de préentraînement massifs, ce qui transfère une partie du coût plutôt qu'il ne l'élimine. La balance économique dépend du rapport entre coût d'annotation médicale et coût compute, lequel évolue rapidement.

6. Enfin, la dérive temporelle. Un encodeur SSL préentraîné en 2025 sur un corpus distributionnellement caractéristique de cette période n'a aucune garantie de rester valide en 2030, lorsque les protocoles d'imagerie, les démographies de cohorte ou les modalités d'acquisition auront évolué. Cette dérive est documentable et requiert un dispositif de surveillance ; elle n'est pas absente du paysage simplement parce que l'encodeur a été préentraîné une fois.

JEPA déplace le problème de l'apprentissage. Il ne le résout pas entièrement.

11. Le seuil non franchi : trois conditions opératoires, et les architectures qui prétendent les satisfaire

Reste la question articulée à la Partie 3/3, et qui constitue le point de vigilance le plus important. Qu'apporte JEPA, ou plus généralement les architectures de world modeling et les architectures agentiques avec mémoire externe, au problème de la mémoire biographique telle que définie en §1 ?

Tenons-nous à la formulation strictement fonctionnelle. Trois conditions opératoires distinguent une mémoire biographique d'une cohérence latente.

1. Première condition : la ***réindexation contextuelle***. Une mémoire biographique permet d'accéder à un contenu mnésique selon plusieurs voies d'entrée (modale, temporelle, affective) et de réactiver à partir de chacune une configuration cohérente de l'épisode entier. Un JEPA produit des représentations stables sous transformation, mais cette stabilité est définie sur un seul axe : la prédictibilité inter-représentationnelle. Il n'y a pas de structure d'indexation multi-perspective dans l'espace latent.

2. Deuxième condition : l'***intégration multi-épisode***. Une mémoire biographique articule des épisodes distincts entre eux par des relations qui ne sont ni purement statistiques ni purement temporelles, mais structurées par une histoire d'agent. Un JEPA apprend des régularités sur l'ensemble du corpus de préentraînement, sans préservation différenciée d'épisodes individuels.

3. Troisième condition : la ***persistance d'agent***. Cette condition demande une définition opératoire, faute de quoi elle reste une formule. Par persistance d'agent, on entend la continuité dans le temps d'un référent unique auquel les arêtes mnésiques sont indexées, ***avec*** la propriété supplémentaire que cet agent peut traiter les épisodes passés comme épisodes vécus par lui-même, et non comme données extérieures consultables. La distinction entre

indexation et appartenance est cruciale. Un journal indexe des événements à un identifiant ; il ne les fait pas appartenir à un sujet.

Une objection sérieuse mérite d'être examinée frontalement. Plusieurs familles d'architectures agentiques contemporaines revendiquent précisément ce que les trois conditions paraissent décrire.

- Les *generative agents* de Park et collaborateurs [35] équipent des LLM d'un ***memory stream*** (flux d'observations indexées), d'un mécanisme de ***reflexion*** (synthèse périodique en méta-souvenirs) et d'un système de récupération combinant similarité, récence et importance.
- Voyager [36] dote un agent Minecraft d'une ***skill library*** persistante et d'un curriculum automatique.
- ReAct [37] et ses extensions intercalent raisonnement, action et mémoire épisodique simple.

Ces architectures sont sérieuses et ne peuvent être écartées par une formule.

Examinons-les à la lumière des trois conditions, sans complaisance :

- Sur la ***réindexation contextuelle***, les ***generative agents*** offrent effectivement une récupération multi-critère (similarité sémantique, récence temporelle, importance pondérée). C'est une forme d'indexation multi-axe, mais qui opère sur des entrées textuelles homogènes ; elle ne réactive pas une configuration multimodale d'épisode, elle compose un prompt à partir de fragments textuels sélectionnés. La distinction est opératoire : une réindexation contextuelle au sens fort réactive l'épisode ; le memory stream le ré-articule. Voyager n'a pas de réindexation au sens épisodique, ses skills sont indexés par fonctionnalité.

- Sur l'***intégration multi-épisode***, les ***generative agents*** ont un mécanisme dédié, la reflexion, qui synthétise périodiquement le memory stream en propositions de plus haut niveau. C'est une intégration, mais elle est compressive et lossy : elle produit des résumés, pas des relations préservant l'individualité des épisodes. Voyager intègre les compétences acquises, pas les épisodes. ReAct n'intègre rien au-delà de la fenêtre courante.

- Sur la ***persistance d'agent***, c'est la condition qui les distingue le plus nettement de la mémoire biographique. Ces architectures ont toutes un identifiant persistant et un journal d'épisodes attaché à cet identifiant. Elles ne satisfont cependant pas la condition d'appartenance : l'agent ne traite pas le memory stream comme des épisodes vécus par lui-même, il le consulte comme une base de données indexée à son identifiant.

La récupération est une opération d'index, pas une réactivation située.

Cette distinction n'est pas de la coquetterie philosophique : elle a des conséquences fonctionnelles vérifiables, notamment sur la capacité à modifier des engagements antérieurs en cohérence avec une trajectoire personnelle plutôt que par sélection de fragments compatibles avec la requête courante.

Aucune de ces architectures ne satisfait donc simultanément les trois conditions au sens strict. Elles s'en approchent, parfois de manière saisissante, mais elles opèrent par juxtaposition : un encodeur, un journal, un mécanisme de récupération, un agent identifié. La question n'est pas de savoir si l'on peut représenter une biographie en juxtaposant ces éléments, mais si la juxtaposition produit la propriété d'appartenance, ou seulement son simulacre fonctionnel utile.

Une précision philosophique s'impose à ce point, pour ne pas laisser planer l'ambiguïté avec la position défendue dans la Partie 3/3 de la série. Cette dernière argumentait que la mémoire biographique humaine se distingue par des propriétés irréductiblement phénoménologiques : autonoéticité au sens de Tulving, modulation affective, navigabilité multi-perspective. La translation en trois conditions opératoires effectuée ici n'abolit pas cette défense. Elle isole un ***minimum opératoire*** en deçà duquel on peut affirmer que le seuil n'est pas franchi, indépendamment de tout engagement phénoménologique. Si une architecture ne satisfait pas la réindexation contextuelle, l'intégration multi-épisode et la persistance d'agent au sens fort, elle ne franchit pas le seuil, quels que soient les arbitrages

métaphysiques sur la conscience. Si elle les satisfait, la question phénoménologique de l'autonoéticité reste ouverte, comme excédent structurel au-dessus du minimum opératoire. Cette stratification permet de tenir une position défendable sans engager la querelle phénoménologique à chaque évaluation architecturale. Elle a un coût : la position phénoménologique forte de la Partie 3/3 cesse d'être nécessaire pour distinguer JEPA d'une mémoire biographique. Elle redevient nécessaire seulement dans la zone où le minimum opératoire serait satisfait, zone que les architectures actuelles n'atteignent pas.

À ce jour, et à ma connaissance, aucune architecture publiée ne satisfait simultanément ces trois conditions opératoires au sens strict défini ici. Ce constat n'est ni une thèse phénoménologique forte, ni un argument d'inaccessibilité de principe. C'est une description architecturale, susceptible d'être révisée par la prochaine publication qui démontrera explicitement la satisfaction des trois conditions, et non leur simulacre par juxtaposition.

12. Conclusion : intelligence et invariance

Ce que les architectures prédictives latentes changent n'est pas la nature de l'intelligence artificielle. C'est la cible de l'apprentissage. Avant elles : apprendre des réponses, classifier, reconstruire. Après elles : apprendre les contraintes qui rendent certaines transformations prédictibles, encoder la cohérence d'un espace plutôt que la fidélité à un signal.

Ce déplacement n'est ni une révolution, ni un détail. Il est un mouvement épistémique précis, dont la portée doit être évaluée à hauteur de ce qu'il fait, à savoir réduire la dépendance aux annotations dans certains régimes, améliorer la robustesse hors distribution sous certaines conditions, structurer la modélisation de trajectoires par contrainte de prédictibilité, et de ce qu'il ne fait pas : satisfaire les conditions fonctionnelles d'une mémoire biographique, ni par lui-même, ni par simple adjonction d'un journal d'épisodes.

JEPA n'est ni une mémoire, ni un simulateur, ni un agent : c'est une architecture qui apprend une géométrie de prédictibilité. Tenir cette formule, c'est accepter de ne pas projeter sur ces systèmes des propriétés qu'ils n'ont pas, et reconnaître celles qu'ils ont effectivement.

La question stratégique pour les architectes industriels qui déploient ces systèmes en environnement régulé n'est donc pas ***faut-il adopter JEPA ?*** La question est faiblement posée. Elle est : ***quelle propriété cherche-t-on à instancier, à quel niveau, et l'architecture choisie l'instancie-t-elle, ou en simule-t-elle seulement la surface ?*** Les deux réponses sont valides selon le contexte, mais elles ne sont pas équivalentes, et leur confusion produit des systèmes qui paraissent intelligents jusqu'au moment exact où on les déplace hors de leur distribution d'entraînement.

L'intelligence ne réside pas dans ce qui est observé, mais dans ce qui ne peut pas varier. Reste à savoir si ce qui ne peut pas varier suffit à constituer un sujet qui se souvient.

Notes et références complémentaires

La numérotation [1] à [26] reprend celle des Parties 1/3, 2/3 et 3/3. Voir : article 1, article 2, article 3

[27] Assran, M. et al. (2023). Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (I-JEPA). CVPR 2023.

[28] Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N. (2024). Revisiting Feature Prediction for Learning Visual Representations from Video (V-JEPA). arXiv:2404.08471.

[29] Chen, T. et al. (2020). A Simple Framework for Contrastive Learning of Visual Representations (SimCLR). ICML 2020.

[30] He, K. et al. (2020). Momentum Contrast for Unsupervised Visual Representation Learning (MoCo). CVPR 2020.

[31] Grill, J.-B. et al. (2020). Bootstrap Your Own Latent (BYOL). NeurIPS 2020.

[32] Caron, M. et al. (2021). Emerging Properties in Self-Supervised Vision Transformers (DINO). ICCV 2021.

[33] He, K. et al. (2022). Masked Autoencoders Are Scalable Vision Learners (MAE). CVPR 2022.

[34] Bardes, A., Ponce, J., LeCun, Y. (2022). VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. ICLR 2022.

[35] Park, J. S. et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. UIST 2023.

[36] Wang, G. et al. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291.

[37] Yao, S. et al. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023.

[38] Azizi, S. et al. (2022). Big Self-Supervised Models Advance Medical Image Classification. Nature Biomedical Engineering.

[39] Zhou, Y., Chia, M. A., Wagner, S. K. et al. (2023). A foundation model for generalizable disease detection from retinal images (RETFound). Nature, 622(7981), 156-163.

[40] Krishnan, R., Rajpurkar, P., Topol, E. J. (2022). Self-Supervised Learning in Medicine and Healthcare. Nature Biomedical Engineering, 6(12).

[41] Tiu, E. et al. (2022). Expert-Level Detection of Pathologies from Unannotated Chest X-Ray Images via Self-Supervised Learning (CheXzero). Nature Biomedical Engineering.

[42] Mei, X. et al. (2022). RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning. Radiology: Artificial Intelligence.

[43] Wang, Z., Sun, J. (2022). SurvTRACE: Transformers for Survival Analysis with Competing Events. ACM-BCB 2022 (arXiv:2110.00855).

Cet article constitue le quatrième volet de la série ***Encodage, transduction et modèles du monde***. Les trois volets précédents (Parties 1/3, 2/3 et 3/3, mars 2026) sont disponibles sur* twingital-ventures.com.
https://twingital-ventures.com/fr/publications/encodage-transduction-modeles-du-monde-1/3 FR
(Summarized English Version available)
https://twingital-ventures.com/fr/publications/encodage-transduction-modeles-du-monde-2/ FR
(Summarized English version available)
https://twingital-ventures.com/fr/publications/encodage-transduction-modeles-du-monde-3/ FR
(Summarized English version available)