Une représentation est une politique de conservation de l'information

Ce que l'apprentissage prédictif latent autorise à inférer sur le monde, et ce qu'il rend irrécupérable en aval

Quatrième volet de la série Encodage, transduction et modèles du monde. Une représentation apprise ne décrit pas un monde : elle décide quelles différences survivent au préentraînement et lesquelles disparaissent. Les architectures prédictives latentes, dont JEPA est l'exemplaire le plus discuté, servent ici de cas d'étude à une question plus générale.

1. Introduction : l'objet réel

Les trois précédents articles de cette série (article 1, article 2, article 3) ont établi que toute architecture cognitive opère par médiation représentationnelle, et que les différences les plus profondes entre systèmes tiennent moins au contenu des représentations qu'à la nature des relations qui les organisent.

Ce volet traite la question qui en découle, et elle n'est pas une question sur JEPA :

Que peut-on légitimement inférer sur le monde à partir d'une représentation apprise, et que devient l'information qu'elle n'a pas conservée ?

La thèse tient en une phrase. ***Un pré-entraînement auto-supervisé est une politique de conservation de l'information : il décide, sous un objectif prétexte, quelles distinctions survivent et lesquelles sont détruites. Les distinctions détruites ne sont pas récupérables par une tête supervisée en aval. Cette irréversibilité, et non la qualité des invariances apprises, est le fait dont dépendent les décisions d'architecture en environnement régulé.***

Quatre non-équivalences soutiennent l'argument, et il faut les tenir ensemble plutôt que séparément :

- **information prédictible ⇏ information stable**
- **information stable ⇏ information causale**
- **information causale ⇏ information décisionnellement pertinente**
- **information décisionnellement pertinente ⇏ information conservée**

La quatrième est celle que la littérature ne traite pas, et c'est celle qui coûte cher.

JEPA sert de cas d'étude parce qu'il rend le mécanisme visible : l'objectif y est explicitement un objectif de prédictibilité, ce qui expose la politique de conservation au lieu de la dissimuler. Les conclusions valent, moyennant adaptation, pour tout pré-entraînement dont l'objectif diffère de la tâche finale.

Le domaine de validité est restreint aux architectures prédictives latentes auto-supervisées et à leur usage en aval par projection supervisée. Les axes de persistance mnésique et de modèle de soi, absents ici, font l'objet du volet suivant, *Une mémoire persistante n'est pas une mémoire biographique*.

2. Trois niveaux à ne pas confondre

1. Le niveau ***computationnel*** concerne la mécanique : encodeurs, espaces latents, objectifs, régularisations, dynamique d'optimisation.
2. Le niveau ***épistémique*** concerne ce qui est appris au sens fort : quelle information est capturée, quelle information est détruite, sous quelles conditions de validité.
3. Le niveau ***pragmatique*** concerne ce que le système permet de décider de manière fiable dans un usage spécifié.

Ces niveaux ne sont pas substituables, et les deux erreurs symétriques sont également communes : sur-attribuer, en lisant une propriété du monde dans une propriété de l'optimisation, et sous-attribuer, en refusant à ces architectures le déplacement épistémique réel qu'elles opèrent.

3. Ce que le dispositif modifie

JEPA s'inscrit dans une lignée continue : SimCLR [1], MoCo [2], BYOL [3] qui montre qu'une prédiction asymétrique évite l'effondrement sans terme contrastif, DINO [4], MAE [5], puis I-JEPA [6] et V-JEPA [7]. Le programme de l'apprentissage de représentations par objectif prédictif est antérieur d'au moins une demi-décennie. JEPA en est une configuration particulière, pas un après. Sa spécificité tient à deux choix : cible située par un signal de position, contexte explicitement masqué plutôt que défini par augmentation.

Le principe [8] s'énonce brièvement. Un encodeur représente le contexte, un encodeur cible représente une cible partielle, un prédicteur conditionné sur la position prédit la seconde à partir de la première, et l'objectif est défini sur la similarité entre prédiction et cible dans l'espace latent.

Il serait commode de résumer cela en disant que l'erreur est simplement mesurée ailleurs, dans l'espace latent plutôt que dans l'espace observationnel. Cette description est trop faible, et elle manque l'essentiel. Passer de $L(X, \hat{X})$ à $L(Z, \hat{Z})$ n'est pas une relocalisation métrique, parce que $Z = f(X)$ est lui-même appris. Le dispositif apprend

simultanément une métrique, une cible et une représentation, et modifie du même coup la relation contexte-cible, les symétries imposées, la dynamique entre encodeur et encodeur cible, et la classe des solutions accessibles à l'optimisation.

JEPA ne déplace pas le thermomètre. Il participe à la définition de ce que le thermomètre appelle une erreur.

La formulation défendable est donc : l'objectif de prédiction est déplacé vers un espace latent appris conjointement, ce qui modifie à la fois la métrique d'erreur et la classe d'informations que l'optimisation est incitée à préserver. C'est cette seconde clause qui porte tout l'article.

Reste une précision sans laquelle l'analyse devient un éloge. Un objectif de prédictibilité latente pris seul converge vers la solution dégénérée où toutes les représentations s'effondrent en un point, le collapse. Les propriétés discutées plus bas n'existent qu'en présence de mécanismes anti-collapse explicites : régularisation par variance et covariance [9], encodeur cible mis à jour par moyenne mobile exponentielle, asymétrie d'architecture. Ce qui est appris est défini par l'objectif combiné à ces contraintes, jamais par l'objectif seul.

Une opposition courante veut par ailleurs que les autoencodeurs soient contraints de tout représenter, bruit compris, là où la prédiction latente sélectionnerait. Elle ne tient pas. Goulot d'étranglement, biais inductif, corruption de l'entrée et structure du décodeur éliminent une masse considérable de détails ; symétriquement, une perte latente ne garantit pas que l'information inutile disparaisse, seulement qu'elle n'est pas contrainte à survivre. La différence est une différence de politique de conservation, pas une différence entre conservation et sélection.

4. Ce que l'optimisation produit

Une discipline lexicale s'impose. Ce qu'un JEPA fait, formellement, est de minimiser une distance entre la prédiction issue du contexte encodé et la représentation d'une cible, sous un protocole donné. « Régularités », « cohérence », « géométrie », « contraintes du monde » sont des interprétations du comportement de cette optimisation, et l'écart entre les deux registres est exactement celui que cet article surveille.

L'énoncé le moins risqué est aussi le moins séduisant :

Un JEPA induit une représentation de ce que son dispositif d'apprentissage parvient à rendre prédictible dans la distribution observée.

Le dispositif comprend six facteurs, et non cinq :

structure apprise = f(D, M, L, A, R, O)

où *D* est la distribution, *M* la politique de masquage, *L* l'objectif, *A* l'architecture, *R* les régularisations et *O* la procédure d'optimisation.

L'inclusion de *O* n'est pas un raffinement : deux entraînements aux cinq premiers facteurs identiques produisent des géométries différentes selon l'initialisation, le schedule, la construction des lots, la dynamique de la moyenne mobile et le point d'arrêt. Si l'argument est que la représentation est une propriété du dispositif et non du monde, alors l'optimisation appartient au dispositif.

Deux corollaires méritent d'être posés explicitement, car ils sont souvent affirmés à l'envers.

L'opposition entre espace de features et espace de cohérence ne tient pas : I-JEPA et V-JEPA sont évalués précisément parce que leurs représentations sont exploitables en aval, et un vecteur latent peut être simultanément une feature, une structure géométrique et un support d'invariance. La qualification de « statique » ne tient pas davantage pour V-JEPA, qui traite des blocs spatio-temporels. Ce qu'il faut dire est plus précis : JEPA ne définit pas d'opérateur explicite de transition d'état itérable permettant de dérouler une trajectoire contrefactuelle, ce qui le distingue d'un modèle dynamique comme DreamerV3 [10] sans lui retirer tout traitement temporel.

Il reste à qualifier les stabilités apprises, et une taxonomie linéaire ne suffit pas, car une invariance possède simultanément une origine et un domaine de validité. Ces deux dimensions sont orthogonales.

Origine de la stabilité	**Tient sur l'entraînement**	**Tient hors échantillon IID**	**Tient en nouvel environnement**	**Tient sous intervention**
Structure du monde	oui	oui	oui	oui
Structure d'échantillonnage	oui	oui	non	non
Artefact instrumental	oui	oui	non	non
Induction algorithmique	oui	variable	non	non

La lecture utile est celle des colonnes. Les quatre origines sont indiscernables dans la première, et les trois premières le restent dans la seconde. Rien dans l'observation IID ne permet de distinguer une contrainte physique d'un artefact d'appareil : la discrimination

n'apparaît qu'à partir de la troisième colonne, et exige donc des environnements multiples plutôt que davantage de données.

Un JEPA n'apprend pas ce qui ne peut pas varier. Il apprend ce que son dispositif rend prédictible, là où il a regardé, sous la contrainte de la manière dont on l'a fait regarder.

Toute revendication de robustesse fondée sur l'invariance apprise doit donc spécifier le dispositif qui l'a produite et la colonne dans laquelle elle a été vérifiée. Ce n'est pas une exigence académique : c'est ce qu'un dossier technique devra contenir pour être défendable.

5. Quatre ensembles d'information

Le cœur du problème n'est ni l'invariance ni la causalité. Il devient visible dès qu'on distingue quatre ensembles.

1. ***I(X)*** : l'information présente dans l'observation.
2. ***I_pred*** : la part de *I(X)* que le dispositif parvient à rendre prédictible depuis le contexte.
3. ***I_Z*** : la part effectivement conservée par la représentation.
4. ***I_Y*** : la part pertinente pour une décision aval donnée.

Le préentraînement pousse *I_Z* vers *I_pred*, puisque rien dans l'objectif n'incite à préserver ce qui ne contribue pas à la prédiction latente. Trois régions en résultent, et elles n'ont pas le même statut.

I_pred ∩ *I_Y* est la zone utile, celle que la littérature mesure. *I_pred* \ *I_Y* est conservée sans valeur décisionnelle : un coût, pas un danger. Reste la troisième :

I_Y \ I_pred : l'information décisionnellement pertinente que rien n'incite à conserver.

C'est là que se produit le dommage, et il possède une propriété que la robustesse hors distribution n'a pas. Il est ***irréversible***. Une tête supervisée ne récupère pas une distinction absente de *I_Z* ; elle ne peut que projeter ce qui a survécu. Un fine-tuning complet peut partiellement restaurer, à condition de disposer d'un signal annoté suffisant pour réapprendre ce que le préentraînement a détruit, ce qui annule précisément l'économie d'annotation invoquée pour justifier le préentraînement.

L'énoncé général est donc :

Une architecture prédictive latente peut éliminer systématiquement une information décisionnellement essentielle, précisément parce que celle-ci est faiblement prédictible depuis son contexte.

En médecine, cette configuration est structurelle plutôt qu'accidentelle : un événement sentinelle, une présentation atypique, une lésion de faible prévalence sont exactement ce que le contexte ne permet pas d'anticiper. Le même mécanisme opère en détection de fraude, en cybersécurité, en maintenance et sur toute classe d'événements dont la valeur décisionnelle tient à leur écart aux régularités dominantes.

Ce qui est bruit pour le problème prétexte est parfois le signal du clinicien.

Une précision s'impose pour ne pas laisser passer un raccourci qui affaiblirait l'argument. Ce n'est pas la rareté qui rend un signal invisible. Un événement rare peut être parfaitement prédictible depuis son contexte, et un événement fréquent peut être difficilement prédictible. Ce qui gouverne la conservation est le ***poids du facteur dans l'objectif espéré*** $E_{x\sim D}[L(x)]$: une caractéristique dont la contribution à la perte moyenne est négligeable ne structure pas la représentation, qu'elle soit prédictible ou non. La rareté n'agit que par ce canal, en pondérant faiblement le facteur dans l'espérance.

Ce qu'il faut donc surveiller n'est pas la fréquence des événements d'intérêt mais leur poids relatif dans l'objectif de préentraînement, quantité qui n'apparaît dans aucune des métriques habituellement rapportées. C'est, en termes d'ingénierie, un problème de ***risque de queue représentationnel***, et il est aujourd'hui non instrumenté.

6. Prédictibilité et causalité

Une structure prédictive peut être parfaitement stable et causalement fausse. Zech et collaborateurs [11] montrent qu'un modèle de détection de pneumonie exploite des marqueurs spécifiques au site, avec effondrement en validation externe. DeGrave, Janizek et Lee [12] montrent que des détecteurs COVID-19 sélectionnent des raccourcis, position du patient, annotations de bord, matériel, plutôt que le signal pathologique. Castro, Walker et Glocker [13] formalisent la question pour l'imagerie.

Le point décisif est que ces raccourcis ne sont pas des instabilités. Ils sont ***remarquablement stables*** dans la distribution d'apprentissage, ce qui les place dans la deuxième ou la troisième ligne du tableau de §4 sans qu'aucune observation IID ne permette de les en distinguer. Un objectif d'invariance ne les élimine donc pas ; il peut les renforcer, un artefact d'appareil étant exactement ce qui varie le moins à l'intérieur d'un site.

Il serait facile d'en conclure que prédictibilité et causalité sont sans rapport. Ce serait remplacer un réductionnisme par son miroir. La formulation exacte comporte deux volets.

- **invariance observationnelle seule ⇏ information causale**
- **invariance + environnements adéquats + hypothèses d'identifiabilité → information causale possible**

C'est précisément le programme de l'invariant causal prediction [14], qui définit l'invariance recherchée comme stabilité du conditionnel sous changement d'environnement et l'exploite pour identifier des relations causales sous hypothèses explicites. L'invariant risk minimization [15] tente de l'opérationnaliser en apprentissage profond, avec un écart bien documenté entre l'intuition et ce que l'objectif garantit [16]. Le causal representation learning [17] pose la question au niveau pertinent : à quelles conditions une représentation apprise correspond-elle à des variables sur lesquelles il est possible d'intervenir ?

Ce qui sépare ces travaux d'un préentraînement auto-supervisé standard n'est donc pas une frontière logique mais un dispositif expérimental. La stabilité observée devient informative sur la structure causale lorsque l'on dispose d'environnements distincts et d'hypothèses d'identifiabilité déclarées. Un JEPA entraîné sur un mélange multi-sites sans traitement explicite de l'environnement ne remplit aucune des deux conditions : il apprend la meilleure géométrie prédictive du mélange, ce qui n'est pas la même chose.

Une représentation peut être invariante et fausse. Elle est alors invariante pour la mauvaise raison, et c'est en déploiement qu'on l'apprend.

7. Un arbitrage, pas un défaut

Dire que l'invariance a un coût suggère un défaut corrigible. Le problème est mieux posé comme un arbitrage entre propriétés incompatibles.

La notion de robustesse hors distribution est d'abord trop grossière pour porter un argument. Elle recouvre des phénomènes hétérogènes appelant des réponses différentes [18] [19] : dérive de covariables, dérive d'étiquettes, dérive de concept, dérive d'acquisition, dérive démographique, dérive temporelle. Sur la dérive de concept en particulier, une représentation qui a stabilisé la relation antérieure est un handicap, pas une protection.

Le mécanisme est explicite. Soit $Z = f(X)$ et supposons que l'objectif encourage $f(X) \approx f(T(X))$ pour une famille de transformations T. L'opération est bénéfique si et seulement si T préserve les variables pertinentes pour la tâche aval, laquelle n'est pas connue au moment du préentraînement. Toute compression optimale l'est relativement à une variable cible [20] ; en son absence, le critère interne de prédictibilité en tient lieu par

défaut, avec les conséquences de §5. Il existe par ailleurs des résultats montrant qu'une invariance représentationnelle poussée peut dégrader la performance cible en présence de dérive d'étiquettes [21], et une caractérisation de l'émergence conjointe d'invariance et de minimalité dans les représentations profondes [22].

Écrire qu'il n'existe pas de représentation universelle optimale serait presque tautologique, l'optimalité supposant un critère. L'énoncé intéressant porte sur la structure de l'arbitrage :

Compression, suffisance, invariance et transférabilité vers des tâches non spécifiées forment un front de Pareto. Aucune représentation ne maximise les quatre.

Une représentation fortement comprimée peut être suffisante pour Y_1 et catastrophique pour Y_2. Une représentation riche préserve les deux au prix de la robustesse, du coût et de l'identifiabilité. La décision d'architecture consiste à choisir un point sur ce front, et la faute industrielle courante consiste à croire qu'on n'en choisit aucun.

8. Géométrie de prédictibilité : un programme d'évaluation

La formule selon laquelle ces architectures apprendraient une ***géométrie de prédictibilité*** est retenue ici avec un statut explicite : celui d'une image, non d'une grandeur. Les observables disponibles ne définissent pas une quantité ordonnable ; elles composent un ***programme d'évaluation multidimensionnel***, ce qui est acceptable à condition de l'appeler ainsi.

Quatre observables sont utilisables, chacune avec sa limite propre.

1. ***Dimension intrinsèque*** de la variété latente. Informative à performance aval égale, mais non critère de qualité en soi : §7 établit qu'une dimension plus élevée peut préserver des facteurs utiles à des tâches non spécifiées.
2. ***Information mutuelle conditionnelle*** entre représentations de contexte et de cible. Estimation notoirement instable en haute dimension continue, et une valeur élevée peut refléter une nuisance partagée. Indice, jamais score.
3. ***Stabilité de la prédiction latente sous perturbations contrôlées***, mesurée par classe et non agrégée. La limite est symétrique de celle du masquage : le résultat dépend du choix des perturbations, c'est-à-dire d'un second dispositif.
4. ***Conservation des facteurs génératifs***, évaluable par sondes sur jeux synthétiques à facteurs contrôlés. Utile en laboratoire, inapplicable en radiologie ou en biologie, où les facteurs génératifs sont précisément ce qu'on ignore.

À cette liste s'ajoute l'observable que §5 rend nécessaire et qui manque partout : la ***mesure de ce qui a été détruit***. Elle est estimable par sondes ciblées sur des distinctions décisionnellement critiques connues, appliquées à la représentation gelée. Un encodeur dont une sonde ne récupère pas une distinction clinique connue est un encodeur qui l'a éliminée, et le constat est plus actionnable que n'importe quel score de robustesse agrégé.

Pour une famille de tâches déclarée et un jeu de perturbations spécifié, on peut comparer deux architectures sur un profil d'observables. On ne peut pas les ordonner sur une grandeur unique.

9. Labels : métrologie plutôt que vérité terrain

Le paradigme dominant présente les annotations comme une ***vérité terrain***. Une taxonomie est plus utile qu'un verdict, à condition qu'elle ne réintroduise pas un réalisme naïf par sa première case.

1. Le label comme résultat de mesure instrumentale : estimation d'une grandeur ou d'un état défini par un protocole métrologique, assortie d'une incertitude, de limites de détection et de conditions de validité. Une mutation BRAF V600E n'est pas un fait brut : le résultat dépend du prélèvement, de l'hétérogénéité tumorale, de la limite de détection, de la technologie, de la qualité de l'échantillon, du pipeline bioinformatique et du seuil d'appel. La métrologie distingue de longue date le mesurande, la mesure et le résultat rapporté [23], et cette distinction manque cruellement au vocabulaire de l'apprentissage supervisé.
2. Le label comme ***construction taxonomique*** : catégories nosologiques dont les frontières résultent d'un consensus révisable.
3. Le label comme ***proxy décisionnel*** : « risque élevé », « répondeur », dont la définition encode un seuil d'action.
4. Le label comme ***convention d'annotation*** : ce que l'annotateur a coché, avec sa variabilité inter-observateur.

Les labels ne sont pas une vérité terrain ; ils sont un résultat de dispositif dont il faut savoir de quel type il relève.

Cette formulation évite la faute symétrique, qui consisterait à traiter tout label comme une pure convention. Un résultat de mesure possède des conditions de validité vérifiables et une incertitude quantifiable, ce qu'une convention d'annotation n'a pas. La différence entre les quatre catégories n'est pas une différence de réalité, elle est une différence de ce qu'il faut documenter pour que le label soit utilisable.

10. Deux canaux, pas deux inductions symétriques

Il est tentant de résumer l'articulation entre auto-supervision et supervision par une symétrie : deux inductions, l'une sous objectif prétexte, l'autre sous objectif d'usage. La formule est commode et masque l'essentiel.

Dans la supervision, le signal d'apprentissage provient partiellement d'une source ***extérieure à X*** : annotation humaine, résultat clinique, mesure biologique, événement futur observé. Dans l'auto-supervision, le signal est construit à partir de la structure de *X* elle-même. Les deux régimes ne diffèrent donc pas par leur objectif, ils diffèrent par leur ***canal d'information***.

auto-supervision : X → contraintes internes de X

supervision : (X, Y) → discrimination, où Y apporte une information absente de X

La conséquence est directe pour §5. Lorsque *Y* apporte une information que *X* ne contient pas, aucun préentraînement, si étendu soit-il, ne peut la produire. Lorsque *Y* dépend d'une information présente dans *X* mais faiblement prédictible depuis son contexte, le préentraînement peut activement la détruire. Le premier cas est une limite de principe, le second est un dommage évitable, et les confondre conduit à surinvestir dans la taille du corpus de préentraînement là où il faudrait investir dans le contrôle de ce qu'il conserve.

L'auto-supervision ne découvre donc pas la structure du domaine. Elle induit des représentations sous un objectif prétexte, et peut induire aussi bien des artefacts, des biais de dataset, des corrélations instrumentales et, dans les cas les plus embarrassants, la structure du pipeline de données lui-même.

11. Dérive endogène : le modèle modifie sa propre distribution

Le traitement usuel de la dérive la suppose exogène : le monde change, la représentation vieillit. En déploiement réel, une partie substantielle de la dérive est produite par le système lui-même.

modèle → décision → monde → nouvelles données → modèle

En santé, un modèle qui oriente le triage modifie la population effectivement examinée, les examens complémentaires prescrits et donc la distribution des images ultérieures. Un modèle qui déclenche une intervention supprime précisément les événements qu'il prédisait, rendant sa propre performance apparente ininterprétable. En crédit et en fraude, la décision détermine quels résultats sont observés, ce qui produit une censure endogène. En maintenance, la prévention supprime les pannes que le modèle apprenait à anticiper.

Ce phénomène est distinct du dataset shift et porte un nom : la prédiction est performative lorsque son émission modifie la distribution qu'elle prédit [24], et le traitement correct exige des modèles contrefactuels du processus de décision plutôt que des modèles prédictifs de l'observation [25].

La conséquence pour un encodeur préentraîné est plus sournoise que le vieillissement. Une représentation optimisée sur la prédictibilité d'un corpus historique encode l'ancienne politique de décision autant que la structure du domaine. Elle stabilise ce que l'ancienne pratique rendait prédictible, y compris les régularités produites par cette pratique, et cette information est indiscernable d'un signal clinique dans toute évaluation IID.

Un système déployé n'observe plus le monde qu'il a appris. Il observe le monde qu'il a contribué à produire.

Une conséquence de gouvernance en découle, rarement formulée : la surveillance après mise sur le marché ne doit pas seulement détecter la dérive, elle doit distinguer la dérive exogène de celle que le déploiement a causée. Les deux appellent des remédiations opposées, réentraînement dans un cas, révision du protocole de décision dans l'autre.

12. Environnement régulé : ce que les preuves autorisent

Le statut logique des preuves disponibles doit être énoncé avant les résultats, car il est systématiquement escamoté. L'essentiel porte sur l'apprentissage auto-supervisé au sens général, pas sur une architecture particulière :

- **Evidence(SSL) ⇏ Evidence(JEPA)**
- **Evidence(SSL) → plausibilité a priori(JEPA)**

Cette distinction interdit de constituer un dossier de validation pour une architecture donnée en s'appuyant sur des résultats obtenus avec d'autres objectifs auto-supervisés, ce qui serait irrecevable en évaluation de conformité.

1. ***Robustesse hors distribution.*** *L'hypothèse d'une moindre dégradation pour un encodeur auto-supervisé préentraîné large a reçu un soutien empirique en imagerie [26] [27] [28]. Elle reste à établir cas par cas, et §7 précise pourquoi : le bénéfice dépend du type de dérive.*

2. ***Dépendance aux datasets annotés.*** *La situation HDLSS pénalise sévèrement la supervision pure, et le préentraînement sur corpus non annotés peut déplacer une partie de la complexité hors de la phase coûteuse en annotation. La revue de Krishnan et collaborateurs [29] documente la promesse et ses conditions restrictives : pour des cohortes spécialisées rares, le corpus de préentraînement n'existe souvent pas dans le domaine cible, et le transfert depuis des référentiels génériques [30] réintroduit des biais que la supervision finale doit traiter.*

3. ***Modélisation de trajectoires.*** *La médecine est essentiellement temporelle, et des architectures apprenant des contraintes de prédictibilité sur paires temporelles peuvent produire des représentations utiles. Nos propres travaux sur trajectoires en oncologie thoracique, conduits avec une architecture d'analyse de survie à événements compétitifs [31], illustrent surtout l'écart en question : une fidélité opérationnelle élevée en aval ne démontre pas que la représentation ait capturé la structure du processus sous-jacent. Le détail méthodologique relève d'une publication distincte.*

Sur le plan réglementaire, un seul principe est mobilisé, et volontairement un seul :

La conformité porte d'abord sur la maîtrise démontrée du système dans son usage prévu ; elle n'exige pas, en règle générale, une interprétation sémantique exhaustive de chaque dimension latente.

Selon la qualification du système, sa classe de risque, la revendication associée et l'analyse de risques, certaines propriétés des sorties intermédiaires ou certaines exigences d'explicabilité peuvent redevenir pertinentes. Ce qui compte pour l'architecte est la ***déployabilité comme propriété système*** : la capacité de l'ensemble, modèle, garde-fous, protocole de validation, dispositif de surveillance et chaîne de décision humaine, à produire des garanties vérifiables. C'est une propriété d'architecture, pas de fonction d'activation.

Une exigence supplémentaire découle de §5, et elle n'est pas couverte par les pratiques actuelles de validation : ***documenter ce que le préentraînement a rendu irrécupérable***. Un dossier technique qui rapporte les performances aval sans caractériser les distinctions détruites décrit ce que le système fait bien, non ce qu'il ne pourra jamais faire.

13. Limites

1. Aucune preuve disponible n'établit qu'un JEPA apprenne une physique du monde. Les démonstrations existantes [6] [7] montrent des stabilités relatives au dispositif au sens de §4.
2. L'évaluation reste mal instrumentée. Les observables de §8 ne sont pas standardisées, et la mesure de l'information détruite n'est pratiquée nulle part.
3. La performance dépend fortement du design des masques et des stratégies de génération de paires, choix inductifs légitimes mais à documenter.
4. Le préentraînement exige un compute substantiel, ce qui transfère une partie du coût d'annotation plutôt qu'il ne l'élimine.

5. L'argument de §5 est structurel mais non quantifié : il indique un mécanisme de destruction d'information, il ne fournit pas d'estimateur de son ampleur pour un couple encodeur-domaine donné. C'est la principale faiblesse de cet article.
6. Le traitement de la dérive endogène en §11 est qualitatif. Les méthodes disponibles supposent un modèle explicite de la politique de décision, rarement disponible en pratique clinique.

14. Conclusion : ce qui a été rendu impossible

Il serait tentant de conclure par une hiérarchie : compression, puis invariance, puis causalité, puis temporalité, puis mémoire, puis agentivité. Cette échelle est fausse. On observe des systèmes dotés de mémoire persistante sans modèle causal, d'agentivité sans modèle de soi, de traitement temporel sans causalité. Ces propriétés ne s'empilent pas, et aucune n'est un scalaire : l'invariance dépend d'une famille de transformations, la compression d'une mesure, la structure causale d'hypothèses d'identifiabilité, la temporalité de ce qu'on entend par temporel.

Ce qui se défend est un espace de propriétés, chacune conditionnelle à un protocole d'évaluation déclaré. Une architecture y occupe une région, jamais un rang, et deux architectures ne sont comparables que sous un protocole commun.

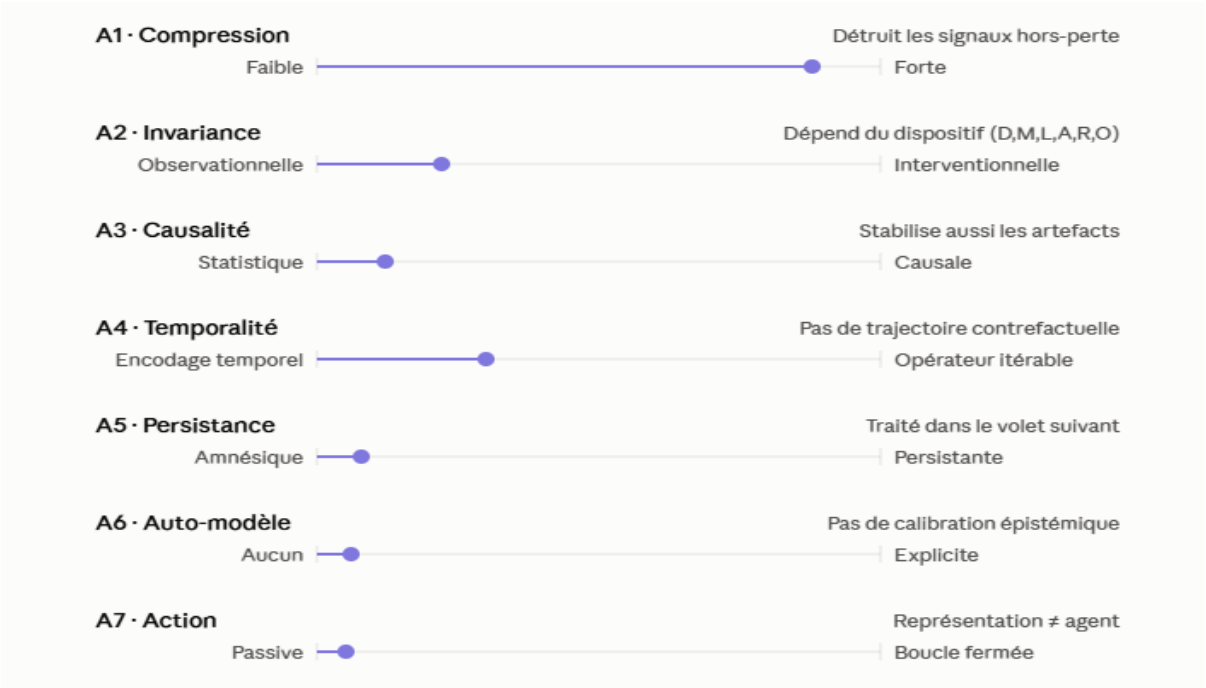

Figure 1. Diagnostic d'architecture : position d'une architecture prédictive latente dans l'espace des propriétés. Chaque axe correspond à une propriété distincte, évaluée séparément. Le curseur indique la position occupée par une architecture de type JEPA au sens du §4 ; l'annotation de droite énonce la contrainte qui détermine cette position. Les axes ne s'empilent pas et ne forment pas une échelle de maturité : on observe des systèmes persistants sans modèle causal, agents sans modèle de soi, temporels sans causalité. Chaque axe est par ailleurs conditionnel à un protocole d'évaluation déclaré, l'invariance

dépendant d'une famille de transformations, la compression d'une mesure, la structure causale d'hypothèses d'identifiabilité. Les positions ne sont comparables ni entre axes, ni entre architectures évaluées sous des protocoles différents. La figure est un diagnostic, pas une mesure.

Le résultat central est ailleurs, et il est plus simple.

Une représentation est une politique de conservation de l'information. Elle conserve certaines différences et en détruit d'autres, et cette destruction est irréversible en aval.

La question de gouvernance qui en découle n'est donc pas celle que l'on pose habituellement. Demander quelles invariances un modèle a apprises revient à inventorier ce qui a survécu. La question utile porte sur le complément :

Quelles distinctions le préentraînement a-t-il rendues irrécupérables en aval, et quelles conséquences cette irréversibilité produit-elle dans l'espace des décisions ?

Elle est répondable, au moins partiellement, par sondage ciblé de la représentation gelée sur des distinctions critiques connues. Elle est aujourd'hui absente des dossiers de validation, des publications d'architecture et des grilles de conformité. C'est précisément le point où l'apprentissage de représentations, la théorie de l'information, la décision sous incertitude et l'ingénierie de la sûreté cessent d'être des disciplines voisines pour devenir le même problème.

L'intelligence des systèmes déployés ne réside ni dans ce qu'ils observent, ni dans ce qu'ils stabilisent. Elle réside dans la justesse des différences qu'ils ont choisi de ne pas détruire. Ce choix n'est pas une propriété du modèle. C'est une décision d'architecte, et elle doit être documentée comme telle.

Les axes de persistance mnésique et de modèle de soi, laissés vides ici, font l'objet du volet suivant : *Une mémoire persistante n'est pas une mémoire biographique*.

Références

1. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning; 2020. p. 1597-1607.

2. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 9729-9738.

3. Grill JB, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, et al. Bootstrap your own latent: a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems. 2020;33:21271-21284.

4. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, et al. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 9650-9660.

5. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 16000-16009.

6. Assran M, Duval Q, Misra I, Bojanowski P, Vincent P, Rabbat M, et al. Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 15619-15629.

7. Bardes A, Garrido Q, Ponce J, Chen X, Rabbat M, LeCun Y, et al. Revisiting feature prediction for learning visual representations from video. arXiv [Preprint]. 2024. arXiv:2404.08471.

8. LeCun Y. A path towards autonomous machine intelligence. OpenReview [Preprint]. 2022;0.9.2.

9. Bardes A, Ponce J, LeCun Y. VICReg: variance-invariance-covariance regularization for self-supervised learning. In: International Conference on Learning Representations; 2022.

10. Hafner D, Pasukonis J, Ba J, Lillicrap T. Mastering diverse domains through world models. arXiv [Preprint]. 2023. arXiv:2301.04104.

11. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15(11):e1002683.

12. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell. 2021;3:610-619.

13. Castro DC, Walker I, Glocker B. Causality matters in medical imaging. Nat Commun. 2020;11:3673.

14. Peters J, Bühlmann P, Meinshausen N. Causal inference by using invariant prediction: identification and confidence intervals. J R Stat Soc Series B Stat Methodol. 2016;78(5):947-1012.

15. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D. Invariant risk minimization. arXiv [Preprint]. 2019. arXiv:1907.02893.

16. Rosenfeld E, Ravikumar P, Risteski A. The risks of invariant risk minimization. In: International Conference on Learning Representations; 2021.

17. Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. Proc IEEE. 2021;109(5):612-634.

18. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors. Dataset shift in machine learning. Cambridge (MA): MIT Press; 2009.

19. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283-286.

20. Tishby N, Zaslavsky N. Deep learning and the information bottleneck principle. In: 2015 IEEE Information Theory Workshop (ITW); 2015. p. 1-5.

21. Zhao H, Des Combes RT, Zhang K, Gordon G. On learning invariant representations for domain adaptation. In: Proceedings of the 36th International Conference on Machine Learning; 2019. p. 7523-7532.

22. Achille A, Soatto S. Emergence of invariance and disentanglement in deep representations. J Mach Learn Res. 2018;19(50):1-34.

23. Joint Committee for Guides in Metrology. International vocabulary of metrology: basic and general concepts and associated terms (VIM). 3rd ed. JCGM 200:2012. Sèvres: BIPM; 2012.

24. Perdomo JC, Zrnic T, Mendler-Dünner C, Hardt M. Performative prediction. In: Proceedings of the 37th International Conference on Machine Learning; 2020. p. 7599-7609.

25. Schulam P, Saria S. Reliable decision support using counterfactual models. In: Advances in Neural Information Processing Systems. 2017;30.

26. Azizi S, Mustafa B, Ryan F, Beaver Z, Freyberg J, Deaton J, et al. Big self-supervised models advance medical image classification. Nat Biomed Eng. 2021;5:1249-1262.

27. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. Nat Biomed Eng. 2022;6:1399-1406.

28. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, et al. A foundation model for generalizable disease detection from retinal images. Nature. 2023;622(7981):156-163.

29. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. Nat Biomed Eng. 2022;6(12):1346-1352.

30. Mei X, Liu Z, Robson PM, Marinelli B, Huang M, Doshi A, et al. RadImageNet: an open radiologic deep learning research dataset for effective transfer learning. Radiol Artif Intell. 2022;4(5):e210315.

31. Wang Z, Sun J. SurvTRACE: transformers for survival analysis with competing events. In: Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; 2022. Article 37.

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