A Representation Is an Information Retention Policy

What latent predictive learning warrants inferring about the world, and what it makes unrecoverable downstream

Fourth installment in the series Encoding, transduction and world models. A learned representation does not describe a world: it decides which differences survive pretraining and which disappear. Latent predictive architectures, of which JEPA is the most discussed exemplar, serve here as the case study for a more general question.

1. Introduction: the real object

The three previous articles in this series (article 1, article 2, article 3) established that every cognitive architecture operates through representational mediation, and that the deepest differences between systems lie less in the content of the representations than in the nature of the relations that organize them.

This installment addresses the question that follows from them, and it is not a question about JEPA:

What can legitimately be inferred about the world from a learned representation, and what becomes of the information it did not retain?

The thesis fits in one sentence. ***Self-supervised pretraining is an information retention policy: under a pretext objective, it decides which distinctions survive and which are destroyed. The destroyed distinctions cannot be recovered by a downstream supervised head. This irreversibility, and not the quality of the learned invariances, is the fact on which architecture decisions depend in regulated environments.***

Four non-equivalences support the argument, and they must be held together rather than separately:

- **predictable information ⇏ stable information**
- **stable information ⇏ causal information**
- **causal information ⇏ decision-relevant information**
- **decision-relevant information ⇏ retained information**

The fourth is the one the literature does not treat, and it is the one that is expensive.

JEPA serves as the case study because it makes the mechanism visible: its objective is explicitly a predictability objective, which exposes the retention policy instead of concealing it. The conclusions hold, with adaptation, for any pretraining whose objective differs from the final task.

The domain of validity is restricted to self-supervised latent predictive architectures and to their downstream use through supervised projection. The axes of memory persistence and self-model, absent here, are the subject of the next installment, *A persistent memory is not a biographical memory*.

2. Three levels not to be confused

1. The ***computational*** level concerns the mechanics: encoders, latent spaces, objectives, regularizations, optimization dynamics.
2. The ***epistemic*** level concerns what is learned in the strong sense: which information is captured, which information is destroyed, under what conditions of validity.
3. The ***pragmatic*** level concerns what the system allows to be decided reliably within a specified use.

These levels are not substitutable, and the two symmetric errors are equally common: over-attribution, by reading a property of the world into a property of the optimization, and under-attribution, by denying these architectures the real epistemic shift they perform.

3. What the apparatus changes

JEPA belongs to a continuous lineage: SimCLR [1], MoCo [2], BYOL [3], which shows that asymmetric prediction avoids collapse without a contrastive term, DINO [4], MAE [5], then I-JEPA [6] and V-JEPA [7]. The program of learning representations under a predictive objective predates it by at least half a decade. JEPA is one particular configuration of that program, not a successor to it. Its specificity rests on two choices: a target located by a position signal, and a context explicitly masked rather than defined by augmentation.

The principle [8] is briefly stated. An encoder represents the context, a target encoder represents a partial target, a predictor conditioned on position predicts the second from the first, and the objective is defined on the similarity between prediction and target in the latent space.

It would be convenient to summarize this by saying that the error is simply measured elsewhere, in latent space rather than in observation space. That description is too weak, and it misses the essential point. Moving from $L(X, \hat{X})$ to $L(Z, \hat{Z})$ is not a metric relocation, because $Z = f(X)$ is itself learned. The apparatus learns a metric, a target and a

representation at once, and thereby modifies the context-target relation, the imposed symmetries, the dynamics between encoder and target encoder, and the class of solutions accessible to the optimization.

JEPA does not move the thermometer. It takes part in defining what the thermometer calls an error.

The defensible formulation is therefore this: the prediction objective is displaced into a jointly learned latent space, which modifies both the error metric and the class of information the optimization is incentivized to preserve. It is this second clause that carries the whole article.

One clarification remains, without which the analysis becomes a eulogy. A latent predictability objective taken alone converges toward the degenerate solution in which all representations collapse to a single point. The properties discussed below exist only in the presence of explicit anti-collapse mechanisms: variance and covariance regularization [9], a target encoder updated by exponential moving average, architectural asymmetry. What is learned is defined by the objective combined with these constraints, never by the objective alone.

A common opposition further holds that autoencoders are constrained to represent everything, noise included, whereas latent prediction would select. It does not hold. Bottleneck, inductive bias, input corruption and decoder structure eliminate a considerable mass of detail; symmetrically, a latent loss does not guarantee that useless information disappears, only that it is not constrained to survive. The difference is a difference of retention policy, not a difference between retention and selection.

4. What the optimization produces

Lexical discipline is called for. What a JEPA does, formally, is minimize a distance between the prediction derived from the encoded context and the representation of a target, under a given protocol. "Regularities", "coherence", "geometry", "constraints of the world" are interpretations of the behavior of that optimization, and the gap between the two registers is exactly the one this article monitors.

The least risky statement is also the least seductive:

A JEPA induces a representation of what its learning apparatus manages to make predictable within the observed distribution.

The apparatus comprises six factors, not five:

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

where *D* is the distribution, *M* the masking policy, *L* the objective, *A* the architecture, *R* the regularizations and *O* the optimization procedure.

The inclusion of *O* is not a refinement: two training runs identical in the first five factors produce different geometries depending on initialization, schedule, batch construction, moving-average dynamics and stopping point. If the argument is that the representation is a property of the apparatus and not of the world, then the optimization belongs to the apparatus.

Two corollaries deserve to be stated explicitly, because they are often asserted backwards.

The opposition between a feature space and a coherence space does not hold: I-JEPA and V-JEPA are evaluated precisely because their representations are exploitable downstream, and a latent vector can be simultaneously a feature, a geometric structure and a support for invariance. The label "static" holds no better for V-JEPA, which processes spatio-temporal blocks. What must be said is more precise: JEPA defines no explicit iterable state-transition operator allowing a counterfactual trajectory to be unrolled, which distinguishes it from a dynamics model such as DreamerV3 [10] without stripping it of all temporal processing.

The learned stabilities remain to be qualified, and a linear taxonomy does not suffice, because an invariance has both an origin and a domain of validity. These two dimensions are orthogonal.

Origin of the stability	**Holds on the training set**	**Holds out of sample, IID**	**Holds in a new environment**	**Holds under intervention**
World structure	yes	yes	yes	yes
Sampling structure	yes	yes	no	no
Instrumental artifact	yes	yes	no	no
Algorithmic induction	yes	variable	no	no

The useful reading is by column. The four origins are indistinguishable in the first, and the first three remain so in the second. Nothing in IID observation makes it possible to distinguish a physical constraint from a device artifact: the discrimination appears only

from the third column onward, and therefore requires multiple environments rather than more data.

A JEPA does not learn what cannot vary. It learns what its apparatus makes predictable, where it was made to look, under the constraint of the way it was made to look.

Any robustness claim founded on a learned invariance must therefore specify the apparatus that produced it and the column in which it was verified. This is not an academic requirement: it is what a technical file will have to contain in order to be defensible.

5. Four sets of information

The core of the problem is neither invariance nor causality. It becomes visible as soon as four sets are distinguished.

1. ***I(X)***: the information present in the observation.
2. ***I_pred***: the share of *I(X)* that the apparatus manages to make predictable from the context.
3. ***I_Z***: the share actually retained by the representation.
4. ***I_Y***: the share relevant to a given downstream decision.

Pretraining pushes *I_Z* toward *I_pred*, since nothing in the objective encourages preserving what does not contribute to latent prediction. Three regions result, and they do not have the same status.

I_pred ∩ *I_Y* is the useful zone, the one the literature measures. *I_pred* \ *I_Y* is retained without decision value: a cost, not a danger. The third remains:

I_Y \ I_pred: the decision-relevant information that nothing encourages retaining.

This is where the damage occurs, and it has a property that out-of-distribution robustness does not have. It is ***irreversible***. A supervised head does not recover a distinction absent from *I_Z*; it can only project what has survived. Full fine-tuning can partially restore it, provided a sufficient annotated signal is available to relearn what pretraining destroyed, which cancels precisely the annotation economy invoked to justify pretraining.

The general statement is therefore:

A latent predictive architecture can systematically eliminate decision-critical information, precisely because that information is weakly predictable from its context.

In medicine, this configuration is structural rather than accidental: a sentinel event, an atypical presentation, a low-prevalence lesion are exactly what the context does not allow one to anticipate. The same mechanism operates in fraud detection, in cybersecurity, in maintenance, and on any class of events whose decision value comes from their departure from the dominant regularities.

What is noise for the pretext problem is sometimes the clinician's signal.

One clarification is required, so as not to let through a shortcut that would weaken the argument. It is not rarity that makes a signal invisible. A rare event can be perfectly predictable from its context, and a frequent event can be hard to predict. What governs retention is the ***weight of the factor in the expected objective*** $E_{x \sim D}[L(x)]$: a feature whose contribution to the average loss is negligible does not structure the representation, whether it is predictable or not. Rarity acts only through that channel, by weighting the factor weakly in the expectation.

What must therefore be monitored is not the frequency of the events of interest but their relative weight in the pretraining objective, a quantity that appears in none of the metrics usually reported. In engineering terms this is a problem of ***representational tail risk***, and it is at present uninstrumented.

6. Predictability and causality

A predictive structure can be perfectly stable and causally wrong. Zech and colleagues [11] show that a pneumonia detection model exploits site-specific markers, with collapse under external validation. DeGrave, Janizek and Lee [12] show that COVID-19 detectors select shortcuts, patient position, edge annotations, hardware, rather than the pathological signal. Castro, Walker and Glocker [13] formalize the question for imaging.

The decisive point is that these shortcuts are not instabilities. They are ***remarkably stable*** within the training distribution, which places them in the second or third row of the table in §4 without any IID observation being able to tell them apart. An invariance objective therefore does not eliminate them; it can reinforce them, a device artifact being exactly what varies least within a site.

It would be easy to conclude that predictability and causality are unrelated. That would replace one reductionism with its mirror image. The exact formulation has two parts.

- **observational invariance alone ⇏ causal information**

- **invariance + adequate environments + identifiability assumptions → causal information possible**

This is precisely the program of invariant causal prediction [14], which defines the sought invariance as stability of the conditional under a change of environment and exploits it to identify causal relations under explicit assumptions. Invariant risk minimization [15] attempts to operationalize it in deep learning, with a well-documented gap between the intuition and what the objective guarantees [16]. Causal representation learning [17] poses the question at the relevant level: under what conditions does a learned representation correspond to variables on which it is possible to intervene?

What separates that body of work from standard self-supervised pretraining is therefore not a logical boundary but an experimental apparatus. Observed stability becomes informative about causal structure when distinct environments and declared identifiability assumptions are available. A JEPA trained on a multi-site mixture without explicit treatment of the environment satisfies neither condition: it learns the best predictive geometry of the mixture, which is not the same thing.

A representation can be invariant and wrong. It is then invariant for the wrong reason, and deployment is where that is learned.

7. A trade-off, not a defect

Saying that invariance has a cost suggests a correctable defect. The problem is better posed as a trade-off between incompatible properties.

The notion of out-of-distribution robustness is, to begin with, too coarse to carry an argument. It covers heterogeneous phenomena calling for different responses [18] [19]: covariate shift, label shift, concept drift, acquisition shift, demographic shift, temporal drift. On concept drift in particular, a representation that has stabilized the earlier relation is a handicap, not a protection.

The mechanism is explicit. Let $Z = f(X)$ and suppose the objective encourages $f(X) \approx f(T(X))$ for a family of transformations T. The operation is beneficial if and only if T preserves the variables relevant to the downstream task, which is not known at pretraining time. Every optimal compression is optimal relative to a target variable [20]; in the absence of one, the internal predictability criterion stands in by default, with the consequences of §5. There are moreover results showing that strong representational invariance can degrade target performance in the presence of label shift [21], and a characterization of the joint emergence of invariance and minimality in deep representations [22].

Writing that there exists no universally optimal representation would be close to tautological, since optimality presupposes a criterion. The interesting statement concerns the structure of the trade-off:

Compression, sufficiency, invariance and transferability to unspecified tasks form a Pareto front. No representation maximizes all four.

A heavily compressed representation can be sufficient for Y_1 and catastrophic for Y_2. A rich representation preserves both at the price of robustness, cost and identifiability. The architecture decision consists in choosing a point on that front, and the common industrial failing consists in believing that no point is being chosen.

8. Predictability geometry: an evaluation program

The formula according to which these architectures learn a ***predictability geometry*** is retained here with an explicit status: that of an image, not of a magnitude. The available observables do not define an orderable quantity; they compose a ***multidimensional evaluation program***, which is acceptable provided it is called that.

Four observables are usable, each with its own limitation.

1. ***Intrinsic dimension*** of the latent manifold. Informative at equal downstream performance, but not a quality criterion in itself: §7 establishes that a higher dimension can preserve factors useful to unspecified tasks.
2. ***Conditional mutual information*** between context and target representations. Notoriously unstable to estimate in high continuous dimension, and a high value can reflect a shared nuisance. An indication, never a score.
3. ***Stability of the latent prediction under controlled perturbations***, measured per class and not aggregated. The limitation is symmetric to that of masking: the result depends on the choice of perturbations, that is, on a second apparatus.
4. ***Retention of the generative factors***, assessable by probes on synthetic datasets with controlled factors. Useful in the laboratory, inapplicable in radiology or biology, where the generative factors are precisely what is unknown.

To this list must be added the observable that §5 makes necessary and that is missing everywhere: the ***measurement of what has been destroyed***. It is estimable by targeted probes on known decision-critical distinctions, applied to the frozen representation. An encoder from which a probe fails to recover a known clinical distinction is an encoder that has eliminated it, and that finding is more actionable than any aggregate robustness score.

For a declared family of tasks and a specified set of perturbations, two architectures can be compared on a profile of observables. They cannot be ordered on a single magnitude.

9. Labels: metrology rather than ground truth

The dominant paradigm presents annotations as ***ground truth***. A taxonomy is more useful than a verdict, provided it does not reintroduce naive realism through its first category.

1. The label as the result of an instrumental measurement: the estimate of a quantity or a state defined by a metrological protocol, together with an uncertainty, detection limits and conditions of validity. A BRAF V600E mutation is not a brute fact: the result depends on the sample, tumor heterogeneity, the detection limit, the technology, sample quality, the bioinformatics pipeline and the calling threshold. Metrology has long distinguished the measurand, the measurement and the reported result [23], and that distinction is sorely missing from the vocabulary of supervised learning.
2. The label as a ***taxonomic construction***: nosological categories whose boundaries result from a revisable consensus.
3. The label as a ***decision proxy***: "high risk", "responder", whose definition encodes an action threshold.
4. The label as an ***annotation convention***: what the annotator ticked, with its inter-observer variability.

Labels are not ground truth; they are the output of an apparatus, and one must know which type of output it is.

This formulation avoids the symmetric error, which would consist in treating every label as a pure convention. A measurement result has verifiable conditions of validity and a quantifiable uncertainty, which an annotation convention does not. The difference between the four categories is not a difference in reality, it is a difference in what has to be documented for the label to be usable.

10. Two channels, not two symmetric inductions

It is tempting to summarize the articulation between self-supervision and supervision by a symmetry: two inductions, one under a pretext objective, the other under a use objective. The formula is convenient and conceals the essential point.

In supervision, the learning signal comes partly from a source ***external to X***: human annotation, clinical outcome, biological measurement, observed future event. In self-supervision, the signal is constructed from the structure of *X* itself. The two regimes therefore differ not in their objective but in their ***information channel***.

self-supervision: X → internal constraints of X

supervision: (X, Y) → discrimination, where Y brings information absent from X

The consequence is direct for §5. When *Y* brings information that *X* does not contain, no pretraining, however extensive, can produce it. When *Y* depends on information present in *X* but weakly predictable from its context, pretraining can actively destroy it. The first case is a limit of principle, the second an avoidable damage, and confusing them leads to overinvesting in the size of the pretraining corpus where the investment should go into controlling what that corpus retains.

Self-supervision therefore does not discover the structure of the domain. It induces representations under a pretext objective, and it can just as well induce artifacts, dataset biases, instrumental correlations and, in the most embarrassing cases, the structure of the data pipeline itself.

11. Endogenous drift: the model modifies its own distribution

The usual treatment of drift assumes it to be exogenous: the world changes, the representation ages. In real deployment, a substantial part of the drift is produced by the system itself.

model → decision → world → new data → model

In healthcare, a model that guides triage modifies the population actually examined, the additional tests ordered and therefore the distribution of subsequent images. A model that triggers an intervention suppresses precisely the events it was predicting, making its own apparent performance uninterpretable. In credit and in fraud, the decision determines which outcomes are observed, which produces endogenous censoring. In maintenance, prevention removes the failures the model was learning to anticipate.

This phenomenon is distinct from dataset shift and has a name: a prediction is performative when its issuance modifies the distribution it predicts [24], and correct treatment requires counterfactual models of the decision process rather than predictive models of the observation [25].

The consequence for a pretrained encoder is more insidious than aging. A representation optimized on the predictability of a historical corpus encodes the former decision policy as much as the structure of the domain. It stabilizes what the former practice made predictable, including the regularities produced by that practice, and this information is indistinguishable from a clinical signal in any IID evaluation.

A deployed system no longer observes the world it learned. It observes the world it helped produce.

A governance consequence follows, rarely formulated: post-market surveillance must not only detect drift, it must distinguish exogenous drift from the drift that deployment

has caused. The two call for opposite remediations, retraining in one case, revision of the decision protocol in the other.

12. Regulated environments: what the evidence permits

The logical status of the available evidence must be stated before the results, because it is systematically glossed over. The essential point concerns self-supervised learning in the general sense, not a particular architecture:

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

This distinction forbids assembling a validation dossier for a given architecture on the basis of results obtained with other self-supervised objectives, which would be inadmissible in a conformity assessment.

1. ***Out-of-distribution robustness.*** *The hypothesis of lesser degradation for a broadly pretrained self-supervised encoder has received empirical support in imaging [26] [27] [28]. It remains to be established case by case, and §7 states why: the benefit depends on the type of drift.*

2. ***Dependence on annotated datasets.*** *The HDLSS situation penalizes pure supervision severely, and pretraining on unannotated corpora can shift part of the complexity out of the annotation-costly phase. The review by Krishnan and colleagues [29] documents the promise and its restrictive conditions: for rare specialized cohorts, the pretraining corpus often does not exist in the target domain, and transfer from generic repositories [30] reintroduces biases that the final supervision has to handle.*

3. ***Trajectory modeling.*** *Medicine is essentially temporal, and architectures learning predictability constraints on temporal pairs can produce useful representations. Our own work on trajectories in thoracic oncology, conducted with a survival analysis architecture for competing events [31], mainly illustrates the gap at issue: high operational fidelity downstream does not demonstrate that the representation has captured the structure of the underlying process. The methodological detail belongs to a separate publication.*

On the regulatory side, a single principle is mobilized, and deliberately only one:

Conformity bears first on the demonstrated control of the system in its intended use; it does not, as a general rule, require an exhaustive semantic interpretation of every latent dimension.

Depending on the qualification of the system, its risk class, the associated claim and the risk analysis, certain properties of intermediate outputs or certain explainability

requirements may become relevant again. What matters for the architect is ***deployability as a system property***: the capacity of the whole, model, guardrails, validation protocol, monitoring apparatus and human decision chain, to produce verifiable guarantees. It is a property of architecture, not of activation function.

A further requirement follows from §5, and it is not covered by current validation practice: ***documenting what pretraining has made unrecoverable***. A technical file that reports downstream performance without characterizing the destroyed distinctions describes what the system does well, not what it will never be able to do.

13. Limitations

1. No available evidence establishes that a JEPA learns a physics of the world. The existing demonstrations [6] [7] show stabilities relative to the apparatus in the sense of §4.
2. Evaluation remains poorly instrumented. The observables of §8 are not standardized, and the measurement of destroyed information is practiced nowhere.
3. Performance depends heavily on mask design and on pair-generation strategies, legitimate inductive choices but ones to be documented.
4. Pretraining requires substantial compute, which transfers part of the annotation cost rather than eliminating it.
5. The argument of §5 is structural but not quantified: it indicates a mechanism of information destruction, it does not provide an estimator of its magnitude for a given encoder-domain pair. This is the main weakness of this article.
6. The treatment of endogenous drift in §11 is qualitative. The available methods assume an explicit model of the decision policy, rarely available in clinical practice.

14. Conclusion: what has been made impossible

It would be tempting to conclude with a hierarchy: compression, then invariance, then causality, then temporality, then memory, then agency. That ladder is false. Systems are observed with persistent memory and no causal model, with agency and no self-model, with temporal processing and no causality. These properties do not stack, and none of them is a scalar: invariance depends on a family of transformations, compression on a measure, causal structure on identifiability assumptions, temporality on what is meant by temporal.

What is defensible is a space of properties, each conditional on a declared evaluation protocol. An architecture occupies a region of that space, never a rank, and two architectures are comparable only under a common protocol.

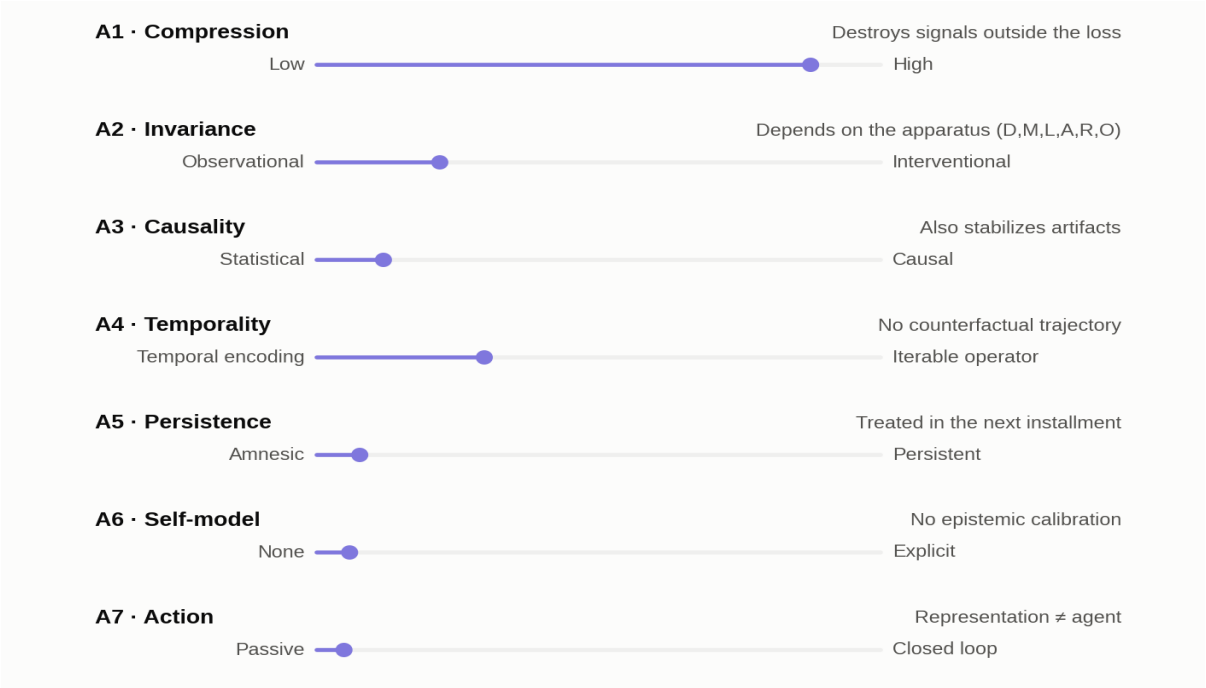

Figure 1. Architecture diagnostic: position of a latent predictive architecture in the space of properties. Each axis corresponds to a distinct property, evaluated separately. The cursor indicates the position occupied by a JEPA-type architecture in the sense of §4; the annotation on the right states the constraint that determines that position. The axes do not stack and do not form a maturity scale: systems are observed that are persistent without a causal model, agentic without a self-model, temporal without causality. Each axis is moreover conditional on a declared evaluation protocol, invariance depending on a family of transformations, compression on a measure, causal structure on identifiability assumptions. Positions are comparable neither across axes nor across architectures evaluated under different protocols. The figure is a diagnostic, not a measurement.

The central result lies elsewhere, and it is simpler.

A representation is an information retention policy. It retains certain differences and destroys others, and that destruction is irreversible downstream.

The governance question that follows is therefore not the one usually asked. Asking which invariances a model has learned amounts to inventorying what has survived. The useful question bears on the complement:

Which distinctions has pretraining made unrecoverable downstream, and what consequences does that irreversibility produce in the space of decisions?

It is answerable, at least in part, by targeted probing of the frozen representation on known critical distinctions. It is today absent from validation dossiers, from architecture publications and from conformity checklists. This is precisely the point at which

representation learning, information theory, decision under uncertainty and safety engineering cease to be neighboring disciplines and become the same problem.

The intelligence of deployed systems lies neither in what they observe nor in what they stabilize. It lies in the aptness of the differences they have chosen not to destroy. That choice is not a property of the model. It is an architect's decision, and it must be documented as such.

The axes of memory persistence and self-model, left empty here, are the subject of the next installment: *A persistent memory is not a biographical memory*.

References

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.

This article is the fourth installment in the series ***Encoding, transduction and world models***. The previous installments are available on twingital-ventures.com.