Certifying Learned Variables

David Freeborn
Northeastern University London
d.freeborn@northeastern.edu

June 23, 2026

Abstract

Machine learning can discover variables that predict the large-scale behavior of physical systems, but prediction alone does not establish that they belong to the system's effective physics. I argue that a learned variable is certified when there is warrant that its governing relationship remains invariant across an independently specified range of irrelevant variations. When the variable is physically opaque, certification must proceed externally, through the learning process or the variable's behavior across that range. The learned Ising coarse-graining can be certified in this way; softness in glass-forming liquids cannot yet. Non-certification, however, does not establish proxyhood. Certification warrants bounded, domain-relative projection, while leaving its stronger realist interpretation open.

1 Introduction

Higher-level physical theories depend on variables that preserve the structure relevant at their scale while suppressing microscopic detail. Temperature, pressure, and magnetization are familiar examples. Physicists usually identify such variables through existing theory, for instance by deriving them from a microscopic model, identifying their symmetries, or showing that they survive coarse-graining. Machine learning offers another route (Freeborn 2025a). Given configurations of the two-dimensional Ising model, a neural network can recover a standard coarse-graining rule and the critical exponent governing behavior near the phase transition (Koch-Janusz and Ringel 2018). Given configurations of a glass-forming liquid, another model learns a per-particle quantity, *softness*, that predicts where the material will rearrange (Cubuk et al. 2015). In both cases, the learned quantity has been presented as more than a useful predictor. It is meant to identify a physical variable of the system.

Predictive success does not establish that conclusion. A quantity can predict accurately while tracking a contingent correlate of the target. A classifier that distinguishes wolves from huskies by detecting snow may perform well until the background changes. The same problem arises in physics. A learned quantity may govern the system's large-scale behavior, or it may be a proxy whose association with that behavior fails under an untested change that should have made no difference. Both possibilities are compatible with success on the observed cases.

The ordinary response is to ask what the candidate variable is. Can it be derived from the microphysics? Does it have the expected symmetry? Does it occupy a known role in an established theory? Learned variables can block this strategy because they are often physically *opaque*.

Their values may be exactly computable even though the physical quantity represented by the computation is unknown. Softness, for example, is an explicit weighted sum of descriptors of a particle's neighborhood. Yet it has not been identified with density, packing geometry, stress, or another independently specified property. We can calculate softness without knowing what, physically, softness is.

This paper separates two questions. The first concerns status: what makes a candidate an effective variable rather than a proxy? The second concerns warrant: what would justify believing that it has that status? The distinction matters because an effective variable may remain uncertified, while a proxy may appear well supported.

My answer to the first question is *invariance*. A candidate is an effective variable when its governing relationship survives the changes that the relevant physical theory treats as irrelevant. These may include changes in microscopic realization, background conditions, or interventions that should leave the higher-level relationship intact. A proxy succeeds over the observed cases but fails under at least one such variation. The distinction is therefore modal. It concerns how the relationship would behave beyond the cases already examined.

My answer to the second question is *certification*. A variable is certified when there is warrant that its governing relationship is invariant across the relevant range. Certification requires both a defensible specification of that range and evidence that the relationship survives across it. The ordering matters. Unless the range is fixed independently of the successes used to test it, there is no non-trivial invariance claim to support.

A universality class supplies the strongest form of range specification. Its members differ microscopically but share the same large-scale behavior because they flow toward the same renormalization-group fixed point. The theory distinguishes relevant perturbations, which alter the large-scale behavior, from irrelevant perturbations, which disappear under coarse-graining. More generally, a successful physical theory may specify a theory-relative range through its mechanism and domain assumptions. In either case, the range must not be defined simply as the set of cases in which the learned relationship happens to work.

The available warrant depends on whether the variable has been physically identified. An *inspection-based* warrant derives the variable from the microphysics, identifies its symmetry, or connects it to a known response relation. An *external* warrant does not require that identity. It appeals instead to the process that produced the variable or to its behavior across an independently specified range. Opacity removes inspection-based routes. Further validation within the same data-generating regime does not replace them, because it establishes success only over a larger observed sample. An opaque variable must therefore be certified externally.

The two cases divide sharply. The Ising model belongs to a well-understood universality class. Renormalization-group theory specifies the irrelevant range; a process-based result connects the learning objective to low-scaling-dimension degrees of freedom; and worked examples recover independently known universal quantities (Gordon et al. 2021; Koch-Janusz and Ringel 2018; Gökmen et al. 2021). These considerations provide an external certificate that does not depend on recognizing the learned map as majority rule. Softness remains physically opaque, and no agreed universality class or comparably developed theory fixes the range across which its barrier relation should remain invariant. It cannot presently be certified.

This negative verdict is limited. Non-certification does not show that softness is a proxy, still less that it is unreal. Its relationship to rearrangement may be invariant across an appropriate range that has not yet been identified. Alternatively, it may fail under a variation independently classified as irrelevant. The present evidence does not decide between those possibilities.

The argument has two further implications. First, prior physical identification is not necessary for warranted variable choice: the magnet shows that physical theory can certify a learned role without first identifying its familiar bearer. Second, certification has bounded reach. It warrants projection across the range fixed by the relevant class or theory, not across relevant changes or outside that domain. The resulting modal commitment can support realism about an effective structure or role, but the certification argument itself does not require that interpretation.

The proposal is narrower than a general theory of scientific variables. It does not claim that invariance is the only consideration relevant to variable choice, that every useful effective theory has a universality class, or that machine learning creates a wholly new epistemic problem. Its contribution is to isolate the additional burden created when a candidate is discovered before its physical identity is available. In that setting, familiar identity-dependent arguments cannot simply be assumed, while prediction remains too weak. Certification names the intermediate achievement and makes its modal requirements explicit.

Section 2 introduces the two cases. Section 3 develops the invariance criterion and the concept of certification. Section 4 shows why opacity leaves only external warrants. Section 5 applies the framework to the magnet and the glass. Section 6 draws the main implications and explains what would be needed to certify softness. Section 7 states a conditional recurrence hypothesis for other domains in which predictive compression may outpace certification.

2 Two learned variables

The argument turns on two cases that initially look similar. In each, a learning procedure produces a low-dimensional quantity from microscopic configurations, the quantity predicts higher-level behavior, and the authors give it a physical interpretation. The cases differ in the physical structure available to support that interpretation. The magnet comes with a mature renormalization-group theory and a well-understood universality class. The glass does not.

2.1 A learned coarse-graining for the magnet

The Ising model represents a magnet as a lattice of spins $s_i \in \{+1, -1\}$. Neighboring spins favor alignment, with energy

$$H(s) = -J \sum_{\langle ij \rangle} s_i s_j,$$

where $J > 0$. At temperature T, the probability of a configuration is proportional to $e^{-H(s)/T}$. At low temperature, aligned configurations dominate; above a critical temperature T_c, thermal fluctuations destroy the order. At criticality, correlated regions occur on every length scale, and the system is scale invariant.

A large-scale description does not track every spin. It replaces groups of spins with coarse variables that retain the information relevant at longer distances. In a block-spin transformation, the lattice is divided into small blocks and each block is assigned an effective spin (Kadanoff 1966). Repeating the transformation removes progressively shorter-range detail.

The renormalization group tracks how the effective Hamiltonian changes under this repeated coarse-graining (Wilson and Kogut 1974). Each step moves the system through a space of possible Hamiltonians. A fixed point is a Hamiltonian whose form is unchanged by further coarse-graining. Near a fixed point, perturbations separate into relevant directions, which grow under repeated rescaling, and irrelevant directions, which contract. If a perturbation is multiplied by b^y when

lengths are rescaled by b, then $y > 0$ marks a relevant direction and $y < 0$ an irrelevant one. Figure 1 depicts this local structure.

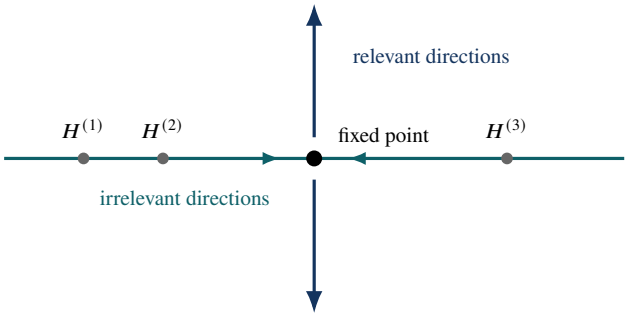

Figure 1: Schematic of a local renormalization-group flow near a critical fixed point. The Hamiltonians $H^{(1)}$, $H^{(2)}$, and $H^{(3)}$ represent microscopically different systems on the same critical surface. They differ only in irrelevant couplings, so repeated coarse-graining suppresses those differences and carries them toward the shared fixed point. A perturbation along a relevant direction instead grows under coarse-graining and moves the system away from the fixed point.

Systems with different microscopic details can flow toward the same fixed point. They then share their critical large-scale behavior and belong to the same universality class. For the two-dimensional Ising class, decisive features include dimensionality, a scalar order parameter with up–down symmetry, and sufficiently short-range equilibrium interactions. Details such as lattice geometry are washed out under coarse-graining. The class therefore supplies a principled distinction between changes that should alter the large-scale physics and changes that should not.

Members of the class share critical exponents. One of these, ν, governs the divergence of the correlation length ξ:

$$\xi \sim |T - T_c|^{-\nu}.$$

The correlation length measures the distance over which spins remain correlated. For the two-dimensional Ising model, $\nu = 1$.

Against this background, Koch-Janusz and Ringel (2018) trained a neural network to discover a coarse-graining from sampled configurations. The network compressed a local block B into a variable h while preserving information about a distant environment E. A buffer separated the block from the environment, preventing the model from succeeding merely by recording short-range details that correlate with nearby spins but disappear at longer distances. Figure 2 shows the geometry of the learning problem and the distinct roles of the block, buffer, and environment. The objective maximized the mutual information

$$I(h; E) = \sum_{h,e} p(h, e) \log \frac{p(h, e)}{p(h)p(e)}.$$

Because the environment lies beyond the buffer, the strongest remaining correlations arise from collective features that persist over long distances. Maximizing $I(h; E)$ therefore encourages the learned variable to retain the degrees of freedom relevant to the large-scale physics.

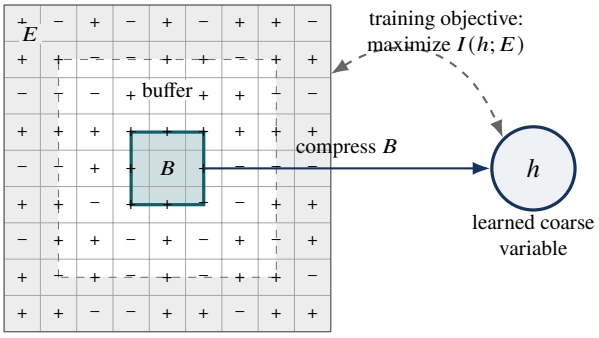

Figure 2: The block–buffer–environment learning setup. A local block B is compressed into a coarse variable h. The intervening buffer screens short-range detail, while the distant environment E supplies the training signal through the mutual information $I(h;E)$. Thus E constrains what information h should retain without being an input to the local coarse-graining map. The diagram is schematic.

For the two-dimensional Ising model, the learned transformation recovers majority-rule coarse-graining: the coarse spin takes the value carried by most spins in the block. Iterating the learned map generates a renormalization-group flow from which the authors estimate

$$\nu \approx 1.0 \pm 0.15,$$

consistent with the exact value. The uncertainty is large by the standards of precision measurement. The result matters because the network was supplied with sampled configurations but no Hamiltonian, prescribed block-spin rule, or target exponent.

The recovery of majority rule might seem less impressive because majority rule is an intuitive coarse-graining for a magnet. A related dimer result answers this concern. A dimer configuration consists of discrete tiles covering adjacent sites on a lattice, while the large-scale theory uses a continuous height field. The learned field cannot simply be a smoothed version of a continuous microscopic input because no such input is present. Applying the same general strategy recovers the emergent field and estimates its scaling dimension as 1.00037, compared with the theoretical value 1 (Gökmen et al. 2021). The method can therefore recover collective variables whose form is not obvious from the microscopic representation.

The authors interpret these results as recovering physically relevant degrees of freedom, not merely as predicting distant configurations. The learned output is presented as a variable of the system's effective physics. The central question is what warrants that stronger interpretation.

2.2 Softness in a glass-forming liquid

The second case concerns a glass-forming liquid. If a liquid is cooled slowly, it may crystallize. If it is cooled quickly enough, crystallization can be avoided. The material remains structurally disordered while its dynamics become progressively slower, until it behaves as a rigid glass on ordinary timescales.

The glass transition is puzzling because the dramatic slowing is accompanied by little visible change in ordinary structure. A snapshot of the liquid and a snapshot of the resulting glass can look remarkably similar even though the time required for particles to move has increased by many orders of magnitude. The slowing is also heterogeneous: most particles remain nearly stationary while small groups undergo sudden local rearrangements. This local predictive question is distinct

from the broader thermodynamic problem of explaining the glass transition.

Cubuk et al. (2015) asked whether a particle's local environment predicts such a rearrangement. Each particle was represented by a vector $\mathbf{F}$ of structure functions describing distances and angular arrangements among its neighbors. Particles were labeled according to whether they were about to rearrange or had remained stable. A support-vector machine was trained to separate the two groups. *Softness* is the signed decision value

$$S = \mathbf{w} \cdot \mathbf{F} - b.$$

Positive values lie on the side associated with rearranging particles and negative values on the side associated with stable ones.

Softness predicts rearrangements with substantial accuracy. In the simulations reported by Schoenholz et al. (2016), roughly ninety percent of particles on the verge of rearranging have positive softness. Perfect prediction is not expected, because thermal fluctuations can trigger or suppress rearrangements even when local structures are similar. The result nevertheless reveals a strong relation between local structure and subsequent motion.

Softness is also related to the rearrangement rate. If it controls a local energy barrier, the probability that a particle of softness S rearranges at temperature T should take the Arrhenius-like form

$$P(\text{rearrange} \mid S) \propto \exp\bigl[\Sigma(S) - \Delta E(S)/T\bigr],$$

where $\Delta E(S)$ is an energy barrier and $\Sigma(S)$ is an entropic contribution. The simulations support this relation, and the inferred barrier varies approximately linearly with softness (Schoenholz et al. 2016). Later work describes softness as a structural order parameter for glassy dynamics, intended to play a role analogous to magnetization in a conventional phase transition (Schoenholz et al. 2017). Hand-designed structural order parameters have been proposed with similar ambitions; one has been described as a "genuine control parameter" of the dynamics (Tong and Tanaka 2019).

Softness is mathematically transparent. Its value is an explicit weighted sum of specified descriptors and can be calculated for any particle. Its physical interpretation remains unsettled. The fitted combination has not been identified with density, packing geometry, mechanical stress, or another independently recognized local property. Exact computability does not decide whether the quantity captures a unified physical feature, combines several such features, or merely tracks a correlation peculiar to the systems and conditions represented in the data.

A second issue concerns governance. No derivation from the underlying physics establishes softness as the variable that controls rearrangement. The evidence shows that softness predicts motion and covaries systematically with an inferred barrier. It does not yet show that the dynamics are organized around softness in the way that the long-distance Ising physics is organized around the relevant scaling fields. A gerrymandered combination and a governing physical variable can agree throughout the observed cases.

Softness is therefore opaque in the sense developed below. The opacity is not computational. The weights and the function they compute are available. The difficulty is physical: inspecting the formula does not reveal which property it represents or establish the role it plays in the dynamics.

More recent graph neural networks predict rearrangements more accurately while learning internal representations that are harder to interpret (Bapst et al. 2020). Their authors generally make more cautious claims, treating the representations as predictors rather than established order parameters. They nevertheless show that the problem is not confined to a simple linear

classifier. As machine learning produces increasingly effective low-dimensional descriptions, the gap between prediction and physical interpretation becomes more rather than less important.

2.3 Scope

The paper concerns learned quantities that satisfy two conditions. First, they are candidates for variables in an effective physical description: coarse fields, order parameters, or effective degrees of freedom rather than labels introduced solely for classification. Second, they are presented as having physical standing rather than merely predictive usefulness. The conditions are independent. A representation may have the right formal shape while its authors remain neutral about its interpretation. The issue here is when a learned candidate can be warranted as belonging to the system's effective physics.

3 Certification, proxies, and invariance

Section 2 introduced two quantities that predict successfully and have the form of variables an effective theory might use. Neither fact settles their physical status. A model can produce a promising coarse quantity without establishing that the quantity belongs to the system's large-scale physics. This section makes that gap precise.

3.1 Prediction, discovery, and certification

Three achievements must be distinguished. By *prediction*, I mean accurate performance on cases that were not used to fit the model. Softness predicts which particles will rearrange; the learned Ising transformation reproduces features of the model's long-distance behavior.

By *discovery*, I mean producing a candidate variable of an effective physical description. The output must have the appropriate form: a coarse or collective quantity of the system, such as an Ising coarse spin or softness, rather than a label introduced solely for classification. Discovery is an achievement even before the candidate's physical standing has been established. It can make a previously unavailable representation available for further inquiry.

By *certification*, I mean possessing warrant that the candidate belongs to the system's effective physics rather than merely fitting the observed cases. Certification is defeasible. Later evidence may reveal that the warrant was unreliable or that the range over which the variable was assessed was wrongly specified.

These achievements can come apart. Prediction shows that a quantity is useful over an observed range. Discovery makes a candidate physical interpretation available. Certification warrants that interpretation. Moving directly from the first two to the third treats predictive success as physical structure already established. Refusing to take a learned candidate seriously until it has been translated into familiar physical terms risks the opposite error: dismissing an important discovery simply because its identity is not yet available (Rice 2025).

The distinction also separates the status of a variable from our epistemic position toward it. A learned quantity may be an effective variable even when we cannot establish that it is. Conversely, a proxy may appear certified until a hidden dependence is exposed. The paper's task is therefore not to define effective variables in terms of our methods for recognizing them, but to state a physical criterion and then ask what evidence can support it.

3.2 The invariance criterion

I propose that a candidate belongs to an effective description when its governing relationship remains stable across every variation that the relevant physical framework treats as irrelevant to that relationship. The criterion is scale- and domain-relative. It does not require the variable to be fundamental, uniquely privileged, or stable under every possible change. It requires the relationship to survive the changes that should leave the effective description intact.

Let s be a many-body system with microstate space Φ_s. A candidate coarse variable is a map

$$V_s : \Phi_s \longrightarrow \mathcal{V}_s,$$

which assigns a coarse value to each microstate, or a coarse field to each local region. For the magnet, V_s is the learned block-spin map. For the glass, it is softness.

Admissible variations may include replacing s with another system. The new system may then have a different microstate space and represent the candidate differently. In such cases, V denotes a role-relative family of maps V_s rather than one literally identical function defined on every system. The issue is whether these representations sustain the same governing role across the relevant changes.

Let Y be the higher-level target that V is proposed to govern. For the magnet, Y is the long-wavelength behavior encoded in the slowest-decaying correlations. For the glass, it is local rearrangement. Let $G[V]$ denote the proposed governing generalization relating V to Y. For softness, this is the barrier relation. For the magnet, it is the claim that the learned field preserves the degrees of freedom controlling the longest-range correlations and their scaling behavior.

A generalization holds at a situation when it correctly describes the dependence of Y on V there. Effective theories rarely hold without qualification, so the criterion permits the corrections and approximations licensed by the theory. Finite-size effects, corrections to scaling, and measurement tolerance do not count as failures when they remain within the stated regime. The estimate $\nu \approx 1.0 \pm 0.15$, for example, counts as agreement with the exact Ising value at the resolution of the learned procedure.

Let $\mathcal{T}$ be a set of admissible variations. A variation may change one of the system's variables, alter its background conditions, or replace it with another system of the same general kind. For $T \subseteq \mathcal{T}$, write

$$\mathrm{Inv}\big(G[V], T\big) \quad \Longleftrightarrow \quad G[V] \text{ holds after every variation } t \in T.$$

The variations divide into those that should preserve the governing relationship and those that may alter it:

$$\mathcal{T} = \mathcal{T}_{\mathrm{irr}} \sqcup \mathcal{T}_{\mathrm{rel}}.$$

The labels are relative to the candidate relationship and the physical domain. Changing the magnet's lattice geometry may be irrelevant to its critical behavior, whereas changing dimensionality or symmetry may move the system into another universality class. Changing temperature or magnetic field may move it along a relevant direction even while the underlying theory continues to describe the resulting flow.

The range may include two conceptually different sorts of change. An intervention alters a variable within the relationship. A background variation changes the circumstances in which the relationship operates, including the microscopic realization of the system. Woodward emphasizes invariance under interventions as central to explanation and treats cross-background persistence as

a form of stability (Woodward 2003, 2010, 2021). The present framework includes both because effective physical variables must often support intervention within a system and transfer across systems.

Let $\mathcal{T}_{\text{obs}} \subseteq \mathcal{T}_{\text{irr}}$ be the irrelevant variations represented in the available evidence. Predictive adequacy supports

$$\text{Inv}\big(G[V], \mathcal{T}_{\text{obs}}\big).$$

Effective-variable status requires invariance across the full irrelevant range.

Definition 3.1 (Effective variable and proxy)**.** Relative to an independently specified physical domain and partition of $\mathcal{T}$, V is an *effective variable* if and only if

$$\text{Inv}\big(G[V], \mathcal{T}_{\text{irr}}\big).$$

It is a *proxy* if and only if this invariance fails: some variation independently classified as irrelevant breaks the governing relationship.

The distinction is modal. A proxy may agree with an effective variable throughout the observed data and diverge only under an untested background change. The familiar wolf classifier illustrates the pattern: detecting snow may classify every observed image correctly while failing when a wolf is photographed indoors or a husky in snow (Ribeiro, Singh, and Guestrin 2016). The omitted background variation exposes the proxy.

The classification is also domain-relative. A quantity can be an effective variable for one restricted theory and a proxy for a broader claim. A local approximation may support invariant prediction within a narrow regime while failing under changes that a more ambitious theory treats as irrelevant. This does not make the criterion arbitrary. The domain and partition must be supplied by independently supported physical commitments, and the status claim is then evaluated relative to them. Disputes about whether a candidate is an effective variable may therefore turn on disputes about the proper effective theory rather than on the candidate's observed accuracy.

The criterion concerns the governing relationship, not constancy of the variable's numerical values or literal identity of its representation. Magnetization changes with temperature and field while remaining the appropriate order parameter. Likewise, a learned map may be reparametrized or implemented differently across systems while preserving the same higher-level dependence. The role expressed by $G[V]$ must survive across the variations the theory regards as irrelevant; its numerical values and implementation need not remain fixed.

Calling $G[V]$ governing does not mean merely that it is unusually stable. The supporting effective framework must treat V as organizing the dependence of Y: changes in V, under the stated background conditions, must support counterfactual expectations about Y. A correlation selected only because it survives the available tests does not meet this condition, since neither its role nor its range has been specified independently. The criterion therefore combines invariance with a theory-supplied candidate role; it is not a purely statistical test for robust association.

The converse is equally important. A quantity may be an effective variable before scientists have tested more than a small part of its domain. Its status depends on how the relationship would behave across $\mathcal{T}_{\text{irr}}$, not on how much of that range has been observed. Additional data from the same source can enlarge $\mathcal{T}_{\text{obs}}$ without closing the modal gap.

3.3 Specifying the irrelevant range

The criterion is informative only if $\mathcal{T}_{\text{irr}}$ is fixed independently of the successes used to test $G[V]$. Otherwise every failure can be classified as relevant after the fact, and every successful variable becomes invariant by definition.

A renormalization-group fixed point supplies the clearest specification.

Definition 3.2 (Class)**.** Fix a renormalization-group fixed point and let $\mathcal{F}$ be a proposed set of defining features, such as dimension, symmetry, and interaction range. The *intension* of the class is

$$\mathcal{S} = \{s : s \text{ instantiates } \mathcal{F}\},$$

while its *extension* is

$$\mathcal{S}^* = \{s : s \text{ flows to the fixed point}\}.$$

Universality is the substantive claim

$$\text{(U)} \qquad \mathcal{S} = \mathcal{S}^*.$$

The intensional side is crucial. If class membership were defined solely by sharing the fixed-point behavior, then projecting that behavior to every class member would be empty. Features such as dimension and symmetry can instead be identified before a candidate system's complete flow is known. The theory then predicts that systems with those features share the fixed point. Because that prediction can fail, (U) has empirical content.

Once the class is established, irrelevant variations are changes that preserve $\mathcal{F}$, remain within the relevant physical regime, and move the system only along directions that contract under coarse-graining. Relevant variations alter a defining feature or excite a direction that grows. The fixed-point framework therefore provides both an intensional range and a structural explanation of why the governing behavior persists across it.

The route by which $\mathcal{F}$ is discovered does not undermine this independence. Physicists may learn which features matter by studying sensitivity near the fixed point. Once identified, however, those features can be specified and checked without rerunning the entire flow or first observing the target behavior in every new system. The requirement is independent specifiability, not metaphysical fundamentality or membership in a natural kind in the strongest sense (Batterman 2021).

Not every effective theory has a useful fixed point. In a *theory-relative* case, a physical theory must specify which changes should preserve the generalization through its mechanism and domain assumptions. This route is less secure but not automatically circular. A theory may state, for example, that a relation should survive changes in density or material composition while failing when a particular interaction or collective mechanism becomes important. If these commitments are fixed before the relevant predictions are assessed, the theory supplies a defeasible but non-trivial range.

The distinction between the two registers concerns the warrant linking the range to the shared behavior. In the fixed-point case, a common flow supplies a structural connection between the defining features and the invariant large-scale behavior. In a theory-relative case, that connection remains inductive. It must be supported by novel predictions, interventions, discriminating comparisons, and transfer across systems. Both registers can support certification, but they do not provide equally strong grounds for projection.

The theory-relative route also raises a circularity problem. Suppose a theory identifies significant variables partly by asking which relationships remain invariant under irrelevant changes, while deciding which changes are irrelevant by asking which preserve those same variables. The criterion would then presuppose the distinction it was introduced to establish. The mechanism and domain assumptions must therefore be supported independently of the learned candidate's predictive record. Fixed-point theory offers the clearest solution because the relevant and irrelevant directions are supplied by an independently developed formalism.

3.4 Certification and its failure modes

Definition 3.3 (Certificate)**.** A learned variable V is *certified* when there is warrant for

$$\mathrm{Inv}\big(G[V], \mathcal{T}_{\mathrm{irr}}\big),$$

where $\mathcal{T}_{\mathrm{irr}}$ is specified independently of the evidence used to establish that invariance.

Certification is defeasible in two ways. First, the physical domain and irrelevant range may be correctly specified while the warrant is unreliable. A learning procedure may produce an apparently stable relationship while tracking the wrong feature, or a limited test may be mistaken for support across the class. This is *warrant-level failure*.

Second, the proposed range may be misindividuated or underdetermined. The defining assumptions may omit a feature on which $G[V]$ depends, so that a variation classified as irrelevant should instead be relevant. Alternatively, distinct partitions may fit the same observed behavior. In either case, evidence can be impeccable relative to the proposed range while the certificate remains defective. This is *range-level failure*.

Range-level assessment is prior. Before asking whether the evidence supports invariance, one must have a defensible account of the range across which invariance is required. A flawless argument over the wrong range cannot certify the variable. This ordering will matter in the glass case, where the principal difficulty arises before the quality of any specific warrant is assessed.

The separation between status and warrant yields four possibilities. An effective variable may be certified, as I will argue for the learned magnet variable. It may also remain uncertified despite belonging to the system's effective physics; this remains possible for softness. A proxy may appear certified when its warrant is defeated. Finally, a proxy may be correctly left uncertified. These possibilities cannot be distinguished by attending only to $\mathcal{T}_{\mathrm{obs}}$.

If the physical domain or partition is not adequately formed, the effective-variable-or-proxy question itself lacks a determinate range. The correct attitude is then suspension, not a positive verdict that the candidate is a proxy. This point prevents an epistemic limitation from being converted into a claim about the world.

Certification also has bounded reach. It warrants projection across $\mathcal{T}_{\mathrm{irr}}$, including unobserved systems and conditions within that range. It does not by itself warrant invariance under relevant variations, outside the specified domain, or relative to another partition. Thus

$$\mathcal{T}_{\mathrm{obs}} \subseteq \mathcal{T}_{\mathrm{irr}} \subseteq \mathcal{T}.$$

Predictive adequacy concerns the first range, certification the second, and neither establishes unrestricted invariance.

4 Opacity and external warrant

Definition 3.3 states what certification requires but not how it can be obtained. Physics supplies several familiar routes to warranted variable choice. Some begin from the physical identity of the variable; others can proceed through its behavior or through the procedure that produced it. Opacity matters because it removes only the first kind.

Definition 4.1 (Opacity)**.** A learned variable V is *opaque*, relative to an epistemic situation, when the available evidence provides no independent physical specification of which quantity it represents, which symmetry it bears, or how it derives from the microscopic description. Full access to the trained model and the function it computes is compatible with opacity in this sense (Humphreys 2009; Creel 2020).

Opacity concerns physical identification, not computational access. Softness is explicit but opaque because its formula does not identify the physical property it represents. A high-dimensional neural representation may be opaque in both senses, but computational complexity is not required. Conversely, a learned variable may cease to be opaque if a later derivation, transformation test, or experiment supplies the missing physical specification.

The definition is relative to the evidence available at a time. This matters because the paper does not divide algorithms into intrinsically interpretable and uninterpretable kinds. The same output can be opaque in one epistemic situation and physically identified in another. The claim developed below is conditional: while the independent specification is unavailable, arguments requiring it cannot be used.

4.1 Inspection-based and external warrants

Consider three representative ways in which physics can warrant a coarse variable. The first uses response relations. The fluctuation–dissipation theorem, for example, connects equilibrium fluctuations with linear response to an applied field and thereby identifies quantities such as susceptibilities and transport coefficients as physically significant (Kubo 1966). Applying the theorem requires the candidate quantity and the perturbation conjugate to it to be specified. Magnetization is paired with an external magnetic field; an uninterpreted model output has no such pairing until its physical role is known.

A second route derives an effective quantity from the microphysics. Homogenization replaces a rapidly varying microstructure with a controlled macroscopic description, as when the elastic properties of a composite are derived from the properties and arrangement of its constituents (Bensoussan, Lions, and Papanicolaou 1978). The derivation identifies both the effective quantity and the assumptions under which it applies. This route also presupposes an account of what variable is being constructed.

A third route appeals to universality. A candidate may be warranted by showing that its governing relationship persists across the irrelevant variations defining a universality class. This argument can concern the learned output's role across the class without first identifying it with a familiar microscopic or macroscopic quantity. A process-based theorem may also show that a learning objective selects the structures independently known to govern the long-distance behavior. The variable's identity may remain unknown while its role is externally constrained.

These examples motivate a general distinction.

Definition 4.2 (Inspection-based and external warrant)**.** A warrant for

$$\mathrm{Inv}\big(G[V], \mathcal{T}_{\mathrm{irr}}\big)$$

is *inspection-based* when its premises require an independent physical specification of V, such as an account of which quantity it represents, which symmetry it bears, or how it derives from the microphysics. It is *external* when they do not.

The distinction concerns the premises of the argument rather than the conventional name of a method. A symmetry inferred by first identifying the output as magnetization is inspection-based. A symmetry established by applying a group transformation to the input and testing how the output changes can be external because it does not presuppose the output's identity. Similarly, a behavioral analogue of a response theorem would count as external if it reached the relevant range without identifying the variable.

External warrants take at least two forms. A *behavioral* warrant appeals to the behavior of V across an independently specified range of systems or conditions. A universality argument is behavioral when it shows that the same governing role persists throughout the class. A *process-based* warrant appeals to the procedure that produced V together with general facts about the physical systems to which the procedure is applied. A theorem connecting an information-bottleneck objective to low-scaling-dimension operators is process-based. The two forms are complementary: process evidence explains why a method should select the right structure, while behavioral calibration checks that it does so in worked cases.

External certification also changes how sameness across systems is understood. A learned variable need not be represented by one literally identical function on every system. Different maps can count as representations of the same effective variable when they sustain the same governing relationship across the relevant variations. A certificate obtained in this way may therefore attach most directly to a role or functional equivalence class. Distinct representations can agree throughout $\mathcal{T}_{\mathrm{irr}}$ while differing elsewhere, so external certification need not settle every fine-grained question about identity.

4.2 The consequence of opacity

Proposition 4.3 (External-warrant requirement)**.** *If a learned variable V is opaque, then no inspection-based warrant is currently available. Validation within the existing data-generating regime establishes at most*

$$\mathrm{Inv}\big(G[V], \mathcal{T}_{\mathrm{obs}}\big).$$

Any available certificate must therefore be external.

Proof. An inspection-based warrant requires an independent physical specification of V, while opacity is the absence of such a specification. Predictive validation on further cases from the same regime may enlarge $\mathcal{T}_{\mathrm{obs}}$, but it does not establish invariance across the full independently specified range $\mathcal{T}_{\mathrm{irr}}$. Any remaining warrant must proceed without identifying V and is therefore external by Definition 4.2. □

The proposition is simple once the categories are fixed, but its consequence is substantive. Opacity does not merely make explanation psychologically difficult. It removes the premises required by derivational and response-based arguments. The surviving route must support a counterfactual projection without first answering the identity question.

Nor is identity-freedom sufficient. Ordinary predictive validation is also identity-free, but it reaches only the observed regime. An external certificate must connect the learned output to an independently specified $\mathcal{T}_{\text{irr}}$, either directly through behavior across that range or indirectly through a reliable process whose relation to that range is established. Without the independent range, external evidence remains a record of successful interpolation or limited transfer.

Process-based evidence must itself be matched to the case. A theorem about an ideal objective does not automatically certify every finite model trained to approximate it. The assumptions connecting the objective to the physical structure must apply, the optimization must produce an adequate solution, and the learned representation must be used within the regime for which the result holds. Behavioral calibration is valuable partly because it checks these links. Conversely, behavioral success without a process account may still certify if it is sufficiently varied and theory-guided, but repeated success over closely related samples offers little protection against a common proxy.

External warrant is therefore not a weaker synonym for testing. It is a family of arguments whose premises do not require prior identification but whose conclusion still concerns the full irrelevant range. Its strength depends on the independence of the range, the diversity and discriminating power of the evidence, and the reliability of the bridge from the learning process to the physical role.

An objection may be raised that an inspection-based warrant can be replaced by a behavioral surrogate. If the surrogate tests $G[V]$ only on the observed distribution or on further cases from the same source, it establishes empirical adequacy rather than certification. If it establishes invariance across $\mathcal{T}_{\text{irr}}$ without invoking an independent specification of V, it is external by Definition 4.2. The surrogate may be scientifically powerful, but it does not reopen the inspection-based route. It instantiates the alternative route that remains.

4.3 Degrees of opacity and strength of warrant

Opacity is often partial in practice. Symbolic distillation may replace a learned representation with a compact expression; transformation tests may reveal a symmetry; feature analysis may connect the output to recognized quantities (Cranmer et al. 2020). Such results can reopen inspection-based routes by supplying part of the missing physical specification. They are not automatically certificates. A recovered formula or symmetry must still support invariance across the independently specified range.

Degrees of opacity should also be distinguished from degrees of warrant. A fully opaque variable may receive a strong fixed-point certificate, while a partly interpreted variable may have only weak predictive support. Among external warrants, a fixed-point certificate has especially secure reach because the class and its irrelevant directions are supplied by a structural theory. A theory-relative certificate may also reach beyond observed cases, but its projection is supported inductively. Bare validation reaches only $\mathcal{T}_{\text{obs}}$.

The account is compatible with computational reliabilism and with views on which opaque models can support scientific understanding through reliable use (Durán and Formanek 2018; Sullivan 2022; Freeborn 2026). It adds a more specific demand. Reliability must support the modal claim relevant to variable status. A process is not enough because it has often predicted correctly. Its reliability must be connected to the variations across which the governing relationship is supposed to persist.

The two cases occupy different positions. The learned Ising transformation is not opaque in

the present epistemic situation because it was identified as majority rule. The certificate developed below is nevertheless external: its force does not depend on that identification. Softness has an explicit formula but lacks a physical specification, while the internal representations of graph networks are more opaque still. In each opaque case, certification requires a range and a warrant that do not presuppose the answer to the identity question.

5 The magnet certified, the glass not

Proposition 4.3 shows that an opaque variable can be certified only externally. Such a certificate still requires a determinate range. The cases differ precisely here. The magnet belongs to a well-understood universality class and receives both process-based and behavioral support relative to that class. Softness remains opaque, and no agreed class or alternative theory fixes the range across which its governing relationship should remain invariant.

5.1 An external certificate for the magnet

Let V be the coarse-graining learned by Koch-Janusz and Ringel (2018), applied across the lattice to produce a coarse field. The relevant domain is the two-dimensional Ising universality class. Its defining features include two spatial dimensions, a scalar order parameter with $\mathbb{Z}_2$ symmetry, sufficiently short-range equilibrium interactions, and the absence of perturbations that carry the system to another fixed point. Microscopically different systems satisfying these conditions, including lattice magnets, lattice gases, and binary alloys, share their critical behavior (Franklin 2018).

The qualification about the basin of attraction matters. Two-dimensionality and up–down symmetry are not by themselves sufficient under arbitrary interactions or disorder. Quenched random fields, sufficiently long-range couplings, or other relevant changes may produce another fixed point or destroy the transition. The class is therefore not an unrestricted resemblance class. It is a physically constrained domain whose boundaries are given by the renormalization-group theory.

Within that domain, reduced temperature and external magnetic field correspond to relevant directions. Changes in lattice geometry, moderate anisotropy, and sufficiently small further-neighbor couplings are irrelevant when they preserve the defining features and remain within the basin of attraction. Renormalization-group theory therefore supplies an independently specified $\mathcal{T}_{\text{irr}}$ and explains why its members share the fixed-point behavior. This is the first component of the certificate: a modal range rather than a list of successful examples.

The governing claim $G[V]$ is that the learned coarse field preserves the low-scaling-dimension degrees of freedom that dominate long-distance correlations. Warrant for its invariance across the class has two further components.

The second is process-based. Gordon et al. (2021) show that, for statistical systems with an appropriate field-theoretic description, information-bottleneck compression retains degrees of freedom associated with the lowest scaling dimensions. These are the operators whose correlations decay most slowly and therefore carry the greatest information between a local block and a distant environment. The objective used in the learned coarse-graining is thus aligned with the structures independently known to control the long-distance behavior. The premises concern the learning objective and the physical class; they do not require the particular learned map to be identified first. Appendix A summarizes the relevant connection.

The third component is behavioral calibration. Iterating the learned Ising transformation yields

$$\nu \approx 1.0 \pm 0.15,$$

consistent with the exact value 1. The result alone would not establish class-wide invariance, since it concerns one model and one universal quantity. Its importance is that it tests the process against a quantity fixed independently by the theory. In the dimer model, the same strategy recovers a non-obvious emergent field with the predicted scaling dimension (Gökmen et al. 2021). The two worked settings differ in microscopic representation and in the form of the large-scale variable. Their agreement with independent theory supports the claim that the objective tracks long-distance structure rather than a peculiarity of one dataset.

The certificate is therefore composite. Renormalization-group theory specifies $\mathcal{T}_{\text{irr}}$ and connects class membership to shared behavior. The process result explains why the learning objective should preserve the appropriate long-range operators. The worked cases calibrate the method against known universal quantities. No component alone is sufficient. Together, they warrant that the learned Ising field sustains its governing role across the irrelevant range.

The learned map was in fact recognized as majority rule. That identification reopens inspection-based routes and provides further reassurance. It is not, however, a premise of the external certificate. The class, process result, and calibration would retain their force if the map were sealed. The magnet is therefore a useful calibration case: because the familiar identity and the universal behavior are independently known, one can check that an identity-free route reaches the right conclusion before relying on such a route where identification is unavailable.

This argument does not require every member of the universality class to be run through the same neural network. The fixed-point theory supplies the projection from the defining features to the shared long-distance structure. The learning results support that the procedure selects the structure to which that projection applies. Demanding exhaustive tests across the entire class would collapse certification back into an impossible enumeration of $\mathcal{T}_{\text{irr}}$ and would ignore the theoretical reason for treating the untested members alike.

Two qualifications remain. First, the process result assumes an appropriate field-theoretic setting, and the strongest direct demonstrations remain a limited set of lattice models. The certificate should not be projected to arbitrary learned coarse-grainings or arbitrary nonequilibrium systems. Second, the conclusion is conditional on the established physical account of Ising universality. These are substantive assumptions, not truths of pure mathematics. Within the domain they define, however, the learned variable is certified as an effective variable.

5.2 Why softness is not yet certified

For softness, Y is local rearrangement and $G[V]$ is the proposed barrier relation

$$P(\text{rearrange} \mid S) \propto \exp\big[\Sigma(S) - \Delta E(S)/T\big].$$

Softness remains opaque. It has not been derived from the microphysics, connected to an established symmetry or conjugate field, or identified with a recognized local structural property. Its explicit formula does not remove this opacity because the physical significance of the fitted combination remains unknown. Inspection-based warrants are therefore unavailable.

An external certificate would require an independently supported account of the materials, regimes, and interventions across which the barrier relation should remain invariant. No such

account is presently agreed. The first difficulty concerns the role of local structure. Reviews describe the connection between structure and dynamical arrest as unresolved (Royall and Williams 2015). Dynamical-facilitation approaches instead explain the slowdown primarily through kinetic constraints and the propagation of mobility, assigning no comparable governing role to a local structural variable (Chandler and Garrahan 2010). This disagreement does not show that softness lacks physical standing. It shows that there is no accepted theory specifying which structural features define the range over which its role should persist.

The second difficulty is the absence of a comparably established fixed-point framework. Competing theories disagree about whether there is a sharp underlying transition, which variables control it, and which systems share the same asymptotic behavior (Berthier and Biroli 2011; Charbonneau et al. 2017). Recurring empirical relaxation laws do not by themselves supply an intensional class. They do not determine which differences among molecular and colloidal glasses, preparation protocols, temperatures, densities, or interaction potentials should count as irrelevant to the softness relation.

The range cannot be recovered simply by surveying the systems in which softness predicts well. Defining the domain as those systems would make the invariance claim retrospective. Nor is it enough to say that the relationship should persist across "similar glass formers." Similarity must be resolved into features that the supporting theory identifies as relevant or irrelevant. At present, no independently supported feature set performs that work.

The result is range-level failure. Without a defensible partition, the claim

$$\mathrm{Inv}\big(G[V], \mathcal{T}_{\mathrm{irr}}\big)$$

has no sufficiently determinate content for the available evidence to support. The obstacle arises before the quality of any particular warrant is assessed. More data can test more systems, but cannot by itself determine which untested changes should have made no difference.

Opacity makes this failure especially consequential. Derivational and response-based arguments can sometimes identify both a variable and the regime in which its relationship holds without requiring a cross-system universality class. A controlled derivation may specify the relevant limiting assumptions; a response relation may identify the quantity, its conjugate perturbation, and the regime of application. Those routes are unavailable while softness lacks a physical specification. The identity-free route remains possible, but it requires the very external range that glass physics has not yet supplied.

Recent work also gives reason for caution at the warrant level. Swain, Ridout, and Nemenman (2024) test the softness procedure in a solvable model whose true barrier is fixed by construction. With clean inputs and unlimited data, the inferred barrier agrees with the true one. With realistic finite data, however, the inferred barrier is systematically too low even though predictive accuracy remains high, the exponential form appears plausible, and the relation looks stable across temperatures. Adding a descriptor correlated with the true cause can also mislead the procedure despite abundant data.

These results expose a precise danger. A fitted quantity can predict well and enter a physically plausible functional relationship while still misidentifying what governs the outcome. Apparent agreement with the barrier form is therefore not automatically evidence that the learned scalar is the barrier-controlling variable. The procedure may exploit a correlated feature or distort the magnitude of the inferred physical quantity while retaining empirical adequacy.

The results do not establish that softness itself is a proxy in the sense of Definition 3.1. The

manipulated feature sets and data conditions test the inference procedure; they are not necessarily physical variations independently classified as irrelevant to actual glassy dynamics. They show warrant-level vulnerability rather than a demonstrated failure of the softness relation under $\mathcal{T}_{\text{irr}}$. This distinction is important. The study weakens the inference from prediction to physical interpretation without deciding the object-level status of softness.

The range-level and warrant-level problems are independent. Even a perfectly reliable learner would need a determinate counterfactual range. Conversely, even if a theory supplied that range, the reliability of the procedure and the adequacy of the evidence would remain further questions. The glass currently faces both obstacles, but the range-level failure is prior.

Two qualifications delimit the conclusion. First, I do not adjudicate among competing theories of the glass transition. The claim concerns the present state of certification, not which theory will ultimately prevail. Second, the laboratory glass temperature T_g is a timescale-dependent crossover at which equilibration becomes impractically slow. Some theories posit a lower ideal transition temperature T_K, but that regime is not accessible under ordinary equilibrium conditions (Berthier and Biroli 2011). These facts explain why a controlled fixed-point certificate is difficult to establish and test; they do not prove that one is impossible.

Softness is therefore uncertified, not refuted. It may be an effective variable for which no adequate certificate yet exists, or it may fail once an appropriate irrelevant variation is identified. The current evidence warrants suspension rather than deflation.

5.3 The bounded reach of certification

The contrast between the magnet and the glass concerns whether a certificate is presently available. A separate question concerns what any certificate establishes once obtained. That question does not depend on whether the warrant is inspection-based or external.

Proposition 5.1 (Bounded reach of certification)**.** *Let V be certified relative to a well-formed partition*

$$\mathcal{T} = \mathcal{T}_{\text{irr}} \sqcup \mathcal{T}_{\text{rel}}.$$

Then the certificate warrants

$$\text{Inv}\big(G[V], \mathcal{T}_{\text{irr}}\big).$$

It does not by itself warrant invariance under variations in $\mathcal{T}_{\text{rel}}$, under variations outside the specified class or theoretical domain, or relative to a different partition of the same variations.

Proof. By Definition 3.3, certification is warrant for invariance across $\mathcal{T}_{\text{irr}}$. Variations in $\mathcal{T}_{\text{rel}}$ are precisely those under which the governing relationship is permitted to change. Nor does a certificate relative to one class or theoretical partition establish invariance relative to another without further argument. □

The proposition separates two constraints. Opacity affects the routes by which certification can be obtained: when the variable lacks an independent physical specification, any available certificate must be external. The partition of variations fixes the certificate's reach: whatever the route, certification warrants invariance across $\mathcal{T}_{\text{irr}}$ and no further without additional argument.

The magnet makes this boundary concrete. Its certificate supports projection across the irrelevant perturbations preserving the Ising class. It does not imply invariance under arbitrary changes in temperature, magnetic field, dimensionality, symmetry, interaction range, or disorder. Some of these changes move the system along relevant directions or into another class. For

softness, the corresponding reach cannot yet be stated because the relevant range has not been fixed. The problem is not that a future certificate would necessarily be narrow, but that no determinate range is presently available over which its breadth could be assessed.

For an opaque variable, the same $\mathcal{T}_{\text{irr}}$ therefore appears twice. It is the range across which an external warrant must work, and it is the boundary of the projectibility thereby secured. These are distinct roles of the same range. The first follows from the loss of identity-dependent warrant; the second follows from the content of the invariance criterion. Identity-indifference does not itself impose the modal boundary, and bounded reach does not follow from opacity.

Certification thus occupies an intermediate modal position:

$$\mathcal{T}_{\text{obs}} \subseteq \mathcal{T}_{\text{irr}} \subseteq \mathcal{T}.$$

Empirical adequacy supports the governing relationship over the observed range. Certification supports projection across the full irrelevant range. Neither establishes unrestricted invariance.

6 Implications and prospects

The framework yields three broader conclusions. First, the epistemology of variable choice must allow certification without prior physical identification. Second, the route proposed for softness is constructive: it specifies what further theory and evidence would have to achieve. Third, certification has a clear modal content while leaving its stronger realist interpretation open.

6.1 Variable choice without prior identification

The magnet places pressure on accounts that require the right variable to be physically identified before its standing can be warranted, and bears on broader disputes about variable choice (Woodward 2016). Batterman's account is a natural comparison because it treats physically significant variables as objective and emphasizes their identification through renormalization-group, homogenization, and response-based reasoning (Batterman 2018, 2021). Nothing in the present argument challenges the objectivity of effective variables. The issue concerns the epistemic order in which their standing can be established. I take Batterman as the clearest developed comparison, not because he explicitly denies the possibility of external certification, but because his account does not explain how physical standing may be warranted when discovery precedes identification.

The learned cases separate two claims that are easily run together. One is metaphysical: the right variables answer to physical structure rather than merely to our convenience. The other is epistemic: warrant for that standing must begin from an independent specification of the variable. The magnet supports the first claim while rejecting the second. Its external certificate depends on the universality class, the learning objective, and independently known universal quantities, not on recognizing majority rule.

One might reply that renormalization-group reasoning already belongs to Batterman's toolkit. The machinery is not new; its epistemic use is. In the familiar identity-first application, renormalization-group reasoning vindicates a variable already specified through its physical character. Here it certifies an invariant role without relying on prior identification of the learned bearer. An account may therefore possess the relevant physical machinery while leaving this identity-independent use unexplained.

The dimer case strengthens the conclusion. The recovered continuous field is not an obvious average of a corresponding continuous microscopic quantity. A procedure can therefore locate a physically relevant role before scientists possess an antecedent description of its bearer. This is not a retreat to pure predictive instrumentalism. The role is certified through independent physical structure, but that structure is connected to the learned output through a different epistemic route.

An adequate epistemology of variable choice should therefore distinguish inspection-based identification from external certification. The first asks what the variable is and uses that answer to justify its role. The second begins from the role, the class, or the reliability of the procedure and can warrant physical standing before a familiar identity is supplied. Later interpretation may add explanatory detail or select among equivalent representations, but it is not always a precondition for certification.

The cases therefore exert two asymmetric pressures. The magnet provides a categorical result: prior physical identification is not necessary for certification. The glass provides a conditional anti-deflationary result: lack of certification does not warrant classifying a candidate as merely predictive. Effective-variable status is a fact about invariance across the physical range; certification is our warrant for that fact. Inferring the absence of the former from the absence of the latter turns an epistemic limitation into a claim about the world. Suspension remains available; deflation requires further evidence.

Three attitudes should therefore be distinguished. A candidate may be certified; judgment may be suspended because the evidence does not determine whether it is an effective variable or a proxy; or the candidate may be deflated and treated as merely predictive. The magnet occupies the first position. Softness presently occupies the second. The third requires positive evidence that the governing relation fails under a variation independently classified as irrelevant. Non-certification alone is not such evidence.

This separation also clarifies the role of availability. Machine learning may make a candidate available before its physical status is known (Rice 2025). Availability concerns how a representation enters inquiry; certification concerns the warrant for treating it as part of the system's physics. An availability-driven discovery may later be externally certified, as in the magnet, or remain useful but uncertified, as in the glass. Data-driven discovery and objective physical standing are compatible, but neither entails the other.

6.2 How softness might be certified

The negative verdict on softness is not an impasse. It identifies two routes by which the present suspension could be overcome. In each, range specification and warrant must remain separate. A theory must first state which variations should preserve the barrier relation; evidence must then support that prediction.

The strongest route would be a controlled fixed-point theory of glassy dynamics. Such a theory would need to identify a set of defining features, an associated universality class, and the relevant and irrelevant directions around the fixed point. This would supply an intensional $\mathcal{T}_{\text{irr}}$ comparable to the one available for the Ising model. Softness would still not be certified merely because the class existed. Its governing relation would have to be shown to track the low-dimensional structure preserved across that class.

There are familiar reasons why this route is difficult. The laboratory transition is an operational crossover, and the putative lower-temperature regimes invoked by some theories are hard or impossible to access in equilibrium. A successful fixed-point account might nevertheless emerge

from analytical work, simulation, new protocols, or a better theory of nonequilibrium scaling. The present criterion is neutral about that possibility. It states what such an account would need to establish.

A second route would use an independently successful dynamical theory in which softness has a specified governing role. The theory would need to state in advance which variations in material, density, temperature, preparation, interaction potential, and spatial organization should preserve the form of the barrier relation. It could not define the range as whatever cases the model successfully fits. Its mechanism and domain assumptions would have to classify the variations before the relevant predictions were tested.

Independent predictions could then provide warrant. Suppose, for example, that the theory predicts that changing density alters the distribution of softness values while preserving the form of the conditional barrier relation. It might derive a consequence for how fragility changes with density. If neither the density dependence nor the consequence was used to construct softness or fit the relation, a successful prediction would test an independent projection beyond $\mathcal{T}_{\text{obs}}$. Similar evidence could come from transfer to new materials, interventions that alter local structure, and discriminating comparisons with rival theories.

The epistemic sequence matters:

1. the theory classifies a variation in advance as belonging to $\mathcal{T}_{\text{irr}}$ for the form of $G[V]$;
2. it derives a consequence of that predicted invariance; and
3. observation or intervention tests the consequence.

The absence of a variation from the construction data does not itself make it irrelevant. Its classification must come from the theory; confirmation then supports that classification rather than defining it.

A single success would rarely suffice. A convincing theory-relative certificate would require a varied body of evidence: quantitative predictions across independently selected variations, transfer to new interaction potentials, stability under reasonable changes in feature construction, intervention on variables the theory classifies as irrelevant, and tests that distinguish the proposed mechanism from facilitation or other rivals. Process-based evidence connecting the learning method to the theory's governing structure would strengthen the case further.

The trap-model program illustrates this strategy. Ridout, Tah, and Liu (2023) construct a model in which each particle's softness determines its energy barrier and rearrangement rate. The model generates consequences beyond the construction of softness, including qualitative predictions concerning density and fragility. Later work introduces interactions by modeling how a rearrangement changes the softness of nearby particles (Ridout and Liu 2024). This is evidence of the right general kind because it embeds softness in a mechanism and derives new consequences.

It is not yet a certificate. The proposed domain, the role of spatial correlations, and the auxiliary variables required alongside softness remain under development. Model revision is scientifically legitimate, but it creates a danger for certification: each failure may be accommodated by changing the mechanism or narrowing the domain. The program must eventually stabilize its commitments enough that some outcomes would count against the proposed range rather than prompting another retrospective redefinition.

Fixed-point and theory-relative certificates would have different strengths. In the first, a shared flow gives a structural connection between the defining features and the common behavior. In the second, the connection remains inductive and defeasible, however extensive the evidence.

Both can certify an effective variable if they specify a non-trivial range and adequately support invariance across it. Effective-variable status is bivalent relative to a fixed domain and partition, while the security and breadth of the warrant for that status are graded.

6.3 The realist significance of certification

Proposition 5.1 fixes the modal content of certification. This content can be stated as follows.

Definition 6.1 (Warranted domain-restricted projectibility)**.** A certificate for V warrants projection of $G[V]$ across the independently specified range $\mathcal{T}_{\text{irr}}$. Without further argument, it does not warrant invariance across $\mathcal{T}_{\text{rel}}$, outside the specified domain, or relative to a different partition of the admissible variations.

This projectibility reaches beyond observed adequacy by licensing expectations about unexamined systems and interventions within $\mathcal{T}_{\text{irr}}$. It is world-directed because the range is fixed by substantive physical claims rather than selected retrospectively, modal because it concerns unobserved variations, and bounded because it extends only as far as the warranted invariance.

A realist may interpret such invariance as evidence for a real scale-relative structure, governing role, or functional equivalence class (Williams 2019; Ladyman and Lorenzetti 2023). The variable need not be fundamental to be real within the effective domain. In a fixed-point case, the boundary is structural rather than merely evidential. Relevant perturbations are amplified under coarse-graining and carry the system away from the fixed point. The renormalization-group argument therefore identifies the directions along which its own invariance conclusion ceases to apply; the same flow structure supports projection across irrelevant directions and marks its limit (Ruetsche 2018).

An empiricist may accept the same counterfactual projection without making that ontological commitment. Certification alone does not establish fundamentality, unrestricted applicability, or a uniquely privileged fine-grained representation. External warrants may identify a role shared by several representations that agree across $\mathcal{T}_{\text{irr}}$ while differing elsewhere. Further evidence would be needed to privilege one bearer (Psillos 1999). The certification argument is therefore detachable from the stronger metaphysical interpretation placed upon it.

7 The predicament recurs

The magnet and glass may look like a contrast between an exceptional failure and the normal case. A more cautious possibility is that the glass predicament recurs in domains where machine learning readily finds low-dimensional predictors but the physical organization needed to certify them is absent. The claim is local and conditional. Fixed points, scaling theories, and well-supported effective variables occur outside equilibrium critical phenomena. The point is only that predictive compression requires less structure than certification.

7.1 Compression without a range of invariance

Sloppiness gives one reason to expect low-dimensional predictive structure to be widespread. In many multiparameter models, the eigenvalues of the Fisher information matrix span many orders of magnitude: predictions are sensitive to a few stiff combinations of parameters and insensitive to many soft combinations (Machta et al. 2013). A high-dimensional model can therefore possess a much lower-dimensional space of distinguishable behavior.

This is parameter-space compression, not itself the construction of a learned effective variable. Sloppiness concerns the map from model parameters to selected predictions, whereas a learned variable $V : \Phi \to \mathcal{V}$ is a function of microstates or configurations. The connection is nevertheless instructive. Both show how successful prediction can depend on far fewer quantities than appear in the underlying description.

Sloppiness is not universality. The two may arise together when coarse-graining drives a system toward a fixed point, rendering many microscopic parameters irrelevant. But a fixed point supplies more than a sensitivity hierarchy: it supplies a flow, an attractor shared by a class of systems, an intensional account of membership, and quantitative invariants. Generic sloppiness supplies none of these. It does not identify a class, an attractor, or a set of variations under which a learned governing relationship should remain invariant (Freeborn 2025b). It can explain why economical prediction is possible without supporting class-wide projectibility.

Sloppiness may also increase evidential underdetermination. Distinct parameterizations or model structures can agree across $\mathcal{T}_{\text{obs}}$ while supporting different projections beyond it. This makes predictive success a weak guide to the physical interpretation of a learned representation and strengthens the need for certification.

No direct inference to proxyhood follows. Soft directions in parameter space are not proxy directions in the space of functions of microstates. For sloppiness to help produce a proxy, a learner would have to select a feature correlated with a governing feature across $\mathcal{T}_{\text{obs}}$, and that correlation would have to break under a variation independently classified as irrelevant. Sloppiness alone establishes none of these claims. It makes certification more important without predetermining its result.

7.2 A conditional recurrence hypothesis

The framework distinguishes three structural positions. A fixed-point theory may supply a formally organized $\mathcal{T}_{\text{irr}}$, as in the magnet. Where no fixed point is available, a mature physical theory may still specify a theory-relative range and support projection through independent predictions, interventions, and transfers. In the third position, neither form of range specification is available. Then the invariance claim cannot yet be assessed, however accurate the learned variable is within $\mathcal{T}_{\text{obs}}$.

The recurrence hypothesis concerns the prevalence of the second and third positions:

> **Recurrence hypothesis.** In domains where low-dimensional predictive structure is common but fixed-point organization is absent, learned variables will commonly lack structurally grounded certification. Unless an independent physical theory specifies and supports a theory-relative range of invariance, they will remain predictive but uncertified.

This is a research hypothesis, not a settled empirical generalization. It predicts neither that most learned variables are proxies nor that certification is impossible in the relevant domains. It predicts that compressed prediction will often arrive before the modal structure needed for certification, and that further validation within the same kind of cases will not supply that structure automatically.

Learned turbulence closures, collective variables in biomolecular simulation, and order parameters in nonequilibrium systems provide natural testbeds. A turbulence closure may perform well within a family of flows while lacking an independently fixed range across Reynolds numbers,

geometries, and forcing conditions. A biomolecular collective variable may accelerate sampling in one molecule without identifying a transferable physical coordinate. An order parameter for active matter may either connect to a controlled symmetry-breaking or scaling theory or remain a suggestive predictor. These are not presumed failures. Each field also contains conservation laws, symmetries, mechanistic constraints, and scaling theories that may furnish inspection-based, fixed-point, or theory-relative warrants.

The practical questions are therefore the same in each case. How was the candidate constructed? Which variations should preserve its governing relationship, and how was that range fixed independently of its predictive record? What evidence supports transfer under those variations? Which failures count against the proposed invariance, rather than prompting a retrospective redefinition of the domain? The glass predicament recurs when these questions lack answers, not merely whenever a learned representation is difficult to interpret.

8 Conclusion

Machine learning can discover candidate variables before their physical identity is understood. This reverses the usual order of inquiry. Rather than beginning from a quantity already specified by theory and asking what it explains, scientists may begin with a representation that predicts successfully and only then ask whether it belongs to the system's physics. The resulting gap is not closed by prediction alone.

The paper has proposed certification as the relevant epistemic achievement. A candidate is an effective variable when its governing relationship remains invariant across the variations that an independently supported physical framework classifies as irrelevant. It is certified when there is warrant for that invariance. Predictive success within the realized regime reaches only $\mathcal{T}_{\text{obs}}$. Certification supports projection across $\mathcal{T}_{\text{irr}}$, including unobserved systems and conditions, but not across relevant changes or outside the domain.

Opacity narrows the routes by which this warrant can be obtained. When the physical identity of a learned variable is unavailable, derivational and response-based arguments lose their premises. Certification must proceed externally through the learning process, the variable's behavior across a specified range, or both. External warrant is not ordinary testing under another name. It must be connected to an independently fixed counterfactual range.

The Ising coarse-graining has such a certificate. Renormalization-group theory fixes the irrelevant range and explains the shared behavior of the class. Process-based results connect the information objective to the long-range operators selected by the theory. The Ising and dimer applications calibrate the method against independently known universal quantities. Recognition of the Ising map as majority rule is welcome but not required by this argument.

Softness does not yet have an equivalent certificate. Its physical identity remains unsettled, no agreed class or mature alternative theory fixes the range across which its barrier relation should persist, and predictive success is compatible with a misleading account of the governing quantity. The proper conclusion is suspension. Softness may be an effective variable for which no adequate warrant is available, or it may be a proxy whose failure will appear under an as-yet-unidentified irrelevant variation.

The same predicament may recur where low-dimensional prediction is available without a fixed point or a mature theory-relative range. Sloppiness and other forms of compression can help explain economical prediction, but they do not by themselves supply a class, an attractor, or a counterfactual partition. This does not show that the resulting variables are proxies. It shows why

they may remain uncertified until further physical structure is supplied.

The right response is therefore neither automatic realism nor automatic deflation. It is to ask which changes should preserve the proposed governing relation, how that range is specified independently, and what evidence supports projection across it. Certification is demanding precisely because it separates successful compression from warranted physical structure.

A The information bottleneck and the transfer-matrix relation

The real-space mutual-information approach of Koch-Janusz and Ringel (2018) learns a coarse variable from a local block while preserving information about its distant environment. Let B denote the microscopic degrees of freedom inside the block, H the learned coarse variable, and E the environment outside an intervening buffer. An information-bottleneck objective can be written as

$$\mathcal{L}_{\mathrm{IB}} = I(B;H) - \beta I(H;E),$$

where $I(B;H)$ penalizes retained microscopic detail and $I(H;E)$ rewards information about the distant environment. Increasing β places greater weight on predictive information about E; the real-space mutual-information limit prioritizes $I(H;E)$ subject to the available compression (Lenggenhager et al. 2020).

Short-distance details are screened from the distant environment, while the degrees of freedom governing long-range correlations remain informative about it. At criticality, these degrees of freedom are associated with operators of low scaling dimension. For a two-dimensional critical system on a cylinder of circumference L, transfer-matrix eigenvalues encode scaling dimensions through

$$\frac{\lambda_i}{\lambda_0} = \exp\left(-\frac{2\pi\Delta_i}{L}\right),$$

up to geometric or anisotropy factors.

Gordon et al. (2021) connect the information-bottleneck solution to this spectrum. Under the assumptions required for the field-theoretic description, the optimal compression retains components associated with the lowest-scaling-dimension operators because these dominate correlations between the block and its distant environment. This supplies the process-based component of the magnet's certificate. The full certificate also requires the Ising universality class to specify $\mathcal{T}_{\mathrm{irr}}$ and the worked Ising and dimer cases to calibrate the procedure against independently known universal quantities.

References

Bapst, Victor, Thomas Keck, Agnieszka Grabska-Barwińska, Craig Donner, Ekin D. Cubuk, Samuel S. Schoenholz, Annette Obika, Alexander W. R. Nelson, Trevor Back, Demis Hassabis, and Pushmeet Kohli. 2020. "Unveiling the predictive power of static structure in glassy systems." *Nature Physics* 16 (4): 448–454. https://doi.org/10.1038/s41567-020-0842-8.

Batterman, Robert W. 2018. "Autonomy of theories: an explanatory problem." *Noûs* 52 (4): 858–873. https://doi.org/10.1111/nous.12191.

Batterman, Robert W. 2021. *A Middle Way: A Non-Fundamental Approach to Many-Body Physics.* New York: Oxford University Press. ISBN: 9780197568613. https://doi.org/10.1093/oso/9780197568613.001.0001.

Bensoussan, Alain, Jacques-Louis Lions, and George Papanicolaou. 1978. *Asymptotic Analysis for Periodic Structures.* Vol. 5. Studies in Mathematics and its Applications. Amsterdam: North-Holland.

Berthier, Ludovic, and Giulio Biroli. 2011. "Theoretical perspective on the glass transition and amorphous materials." *Reviews of Modern Physics* 83 (2): 587–645. https://doi.org/10.1103/RevModPhys.83.587.

Chandler, David, and Juan P. Garrahan. 2010. "Dynamics on the way to forming glass: bubbles in space-time." *Annual Review of Physical Chemistry* 61:191–217. https://doi.org/10.1146/annurev.physchem.040808.090405.

Charbonneau, Patrick, Jorge Kurchan, Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi. 2017. "Glass and jamming transitions: from exact results to finite-dimensional descriptions." *Annual Review of Condensed Matter Physics* 8:265–288. https://doi.org/10.1146/annurev-conmatphys-031016-025334.

Cranmer, Miles, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. 2020. "Discovering symbolic models from deep learning with inductive biases." In *Advances in Neural Information Processing Systems,* 33:17429–17442. Curran Associates, Inc.

Creel, Kathleen A. 2020. "Transparency in complex computational systems." *Philosophy of Science* 87 (4): 568–589. https://doi.org/10.1086/709729.

Cubuk, Ekin D., Samuel S. Schoenholz, Jennifer M. Rieser, Brad D. Malone, Jörg Rottler, Douglas J. Durian, Efthimios Kaxiras, and Andrea J. Liu. 2015. "Identifying structural flow defects in disordered solids using machine-learning methods." *Physical Review Letters* 114 (10): 108001. https://doi.org/10.1103/PhysRevLett.114.108001.

Durán, Juan M., and Nico Formanek. 2018. "Grounds for trust: essential epistemic opacity and computational reliabilism." *Minds and Machines* 28 (4): 645–666. https://doi.org/10.1007/s11023-018-9481-6.

Franklin, Alexander. 2018. "On the renormalization group explanation of universality." *Philosophy of Science* 85 (2): 225–248. https://doi.org/10.1086/696812.

Freeborn, David Peter Wallis. 2025a. "Effective theory building and manifold learning." *Synthese* 205 (1): 23. https://doi.org/10.1007/s11229-024-04844-0.

Freeborn, David Peter Wallis. 2025b. "Sloppy models, renormalization group realism, and the success of science." *Erkenntnis* 90 (2): 645–673. https://doi.org/10.1007/s10670-023-00728-w.

Freeborn, David Peter Wallis. 2026. *A model of understanding in deep learning systems.* https://doi.org/10.48550/arXiv.2604.04171. arXiv: 2604.04171 [cs.AI]. https://arxiv.org/abs/2604.04171.

Gökmen, Doruk Efe, Zohar Ringel, Sebastian D. Huber, and Maciej Koch-Janusz. 2021. "Statistical physics through the lens of real-space mutual information." *Physical Review Letters* 127 (24): 240603. https://doi.org/10.1103/PhysRevLett.127.240603.

Gordon, Amit, Aditya Banerjee, Maciej Koch-Janusz, and Zohar Ringel. 2021. "Relevance in the renormalization group and in information theory." *Physical Review Letters* 126 (24): 240601. https://doi.org/10.1103/PhysRevLett.126.240601.

Humphreys, Paul. 2009. "The philosophical novelty of computer simulation methods." *Synthese* 169 (3): 615–626. https://doi.org/10.1007/s11229-008-9435-2.

Kadanoff, Leo P. 1966. "Scaling laws for Ising models near T_c." *Physics Physique Fizika* 2 (6): 263–272. https://doi.org/10.1103/PhysicsPhysiqueFizika.2.263.

Koch-Janusz, Maciej, and Zohar Ringel. 2018. "Mutual information, neural networks and the renormalization group." *Nature Physics* 14 (6): 578–582. https://doi.org/10.1038/s41567-018-0081-4.

Kubo, Ryōgo. 1966. "The fluctuation–dissipation theorem." *Reports on Progress in Physics* 29 (1): 255–284. https://doi.org/10.1088/0034-4885/29/1/306.

Ladyman, James, and Lorenzo Lorenzetti. 2023. "Effective ontic structural realism." Advance online publication, *The British Journal for the Philosophy of Science,* https://doi.org/10.1086/729061.

Lenggenhager, Patrick M., Doruk Efe Gökmen, Zohar Ringel, Sebastian D. Huber, and Maciej Koch-Janusz. 2020. "Optimal renormalization group transformation from information theory." *Physical Review X* 10 (1): 011037. https://doi.org/10.1103/PhysRevX.10.011037.

Machta, Benjamin B., Ricky Chachra, Mark K. Transtrum, and James P. Sethna. 2013. "Parameter space compression underlies emergent theories and predictive models." *Science* 342 (6158): 604–607. https://doi.org/10.1126/science.1238723.

Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth.* London and New York: Routledge.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. ""Why should I trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1135–1144. https://doi.org/10.1145/2939672.2939778.

Rice, Collin. 2025. "A model-based approach to the problem of variable choice." *Philosophy of Science* 92 (2): 432–452. https://doi.org/10.1017/psa.2024.52.

Ridout, Sean A., and Andrea J. Liu. 2024. *The dynamics of machine-learned "softness" in supercooled liquids describe dynamical heterogeneity.* https://doi.org/10.48550/arXiv.2406.05868. arXiv: 2406.05868 [cond-mat.soft]. https://arxiv.org/abs/2406.05868.

Ridout, Sean A., Indrajit Tah, and Andrea J. Liu. 2023. "Building a "trap model" of glassy dynamics from a local structural predictor of rearrangements." *EPL (Europhysics Letters)* 144 (4): 47001. https://doi.org/10.1209/0295-5075/ad0c70.

Royall, C. Patrick, and Stephen R. Williams. 2015. "The role of local structure in dynamical arrest." *Physics Reports* 560:1–75. https://doi.org/10.1016/j.physrep.2014.11.004.

Ruetsche, Laura. 2018. "Renormalization group realism: the ascent of pessimism." *Philosophy of Science* 85 (5): 1176–1189. https://doi.org/10.1086/699719.

Schoenholz, Samuel S., Ekin D. Cubuk, Efthimios Kaxiras, and Andrea J. Liu. 2017. "Relationship between local structure and relaxation in out-of-equilibrium glassy systems." *Proceedings of the National Academy of Sciences* 114 (2): 263–267. https://doi.org/10.1073/pnas.1610204114.

Schoenholz, Samuel S., Ekin D. Cubuk, Daniel M. Sussman, Efthimios Kaxiras, and Andrea J. Liu. 2016. "A structural approach to relaxation in glassy liquids." *Nature Physics* 12 (5): 469–471. https://doi.org/10.1038/nphys3644.

Sullivan, Emily. 2022. "Understanding from machine learning models." *The British Journal for the Philosophy of Science* 73 (1): 109–133. https://doi.org/10.1093/bjps/axz035.

Swain, Arabind, Sean Alexander Ridout, and Ilya Nemenman. 2024. "Machine learning that predicts well may not learn the correct physical descriptions of glassy systems." *Physical Review Research* 6 (3): 033091. https://doi.org/10.1103/PhysRevResearch.6.033091.

Tong, Hua, and Hajime Tanaka. 2019. "Structural order as a genuine control parameter of dynamics in simple glass formers." *Nature Communications* 10:5596. https://doi.org/10.1038/s41467-019-13606-3.

Williams, Porter. 2019. "Scientific realism made effective." *The British Journal for the Philosophy of Science* 70 (1): 209–237. https://doi.org/10.1093/bjps/axx043.

Wilson, Kenneth G., and John B. Kogut. 1974. "The renormalization group and the ϵ expansion." *Physics Reports* 12 (2): 75–200. https://doi.org/10.1016/0370-1573(74)90023-4.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation.* New York: Oxford University Press. https://doi.org/10.1093/0195155270.001.0001.

Woodward, James. 2010. "Causation in biology: stability, specificity, and the choice of levels of explanation." *Biology & Philosophy* 25 (3): 287–318. https://doi.org/10.1007/s10539-010-9200-z.

Woodward, James. 2016. "The problem of variable choice." *Synthese* 193 (4): 1047–1072. https://doi.org/10.1007/s11229-015-0810-5.

Woodward, James. 2021. "Explanatory autonomy: the role of proportionality, stability, and conditional irrelevance." *Synthese* 198 (1): 237–265. https://doi.org/10.1007/s11229-018-01998-6.