Viewpoint

What Kind of Claims Are Transparency, Explainability, and Interpretability? A Definitional Taxonomy for Health AI

Tessa Ringer[1,2,3], MD, JD, MPhil, CCFP(EM)

[1]Schwartz/Reisman Emergency Centre, Sinai Health System, Toronto, ON, Canada

[2]Division of Emergency Medicine, Department of Family and Community Medicine, University of Toronto, Toronto, ON, Canada

[3]Emergency Department, Peterborough Regional Health Centre, Peterborough, ON, Canada

Corresponding Author:
Tessa Ringer, MD, JD, MPhil, CCFP(EM)
Schwartz/Reisman Emergency Centre
Sinai Health System
Mount Sinai Hospital
206 - 600 University Avenue
Toronto, ON, M5G 1X5
Canada
Phone: 1 416 586 5133
Email: ringer@edmd.ca

Abstract

Transparency, explainability, and interpretability are ubiquitous in the health AI literature, yet are used inconsistently and often interchangeably. This viewpoint does not attempt to define the 3 terms. It asks a prior, meta-level question: what kind of claim do we make when we call a system transparent, explainable, or interpretable? I argue that the 3 terms pick out 3 different kinds of claim. A transparency claim is structural: it concerns what can be inspected about a system. An explainability claim is relational: it concerns whether understanding is successfully conveyed to a specified class of observer. An interpretability claim is functional: it concerns whether a bounded human agent can, in principle, rederive the system's output by following its procedure. I use the framing example of a clinical decision support system that uses machine learning to identify patients at risk of deterioration. I develop each claim type through competing technical accounts, consider salient objections to each, and defend the taxonomy against them. I close by drawing out what the taxonomy implies for the design, evaluation, reporting, and regulation of health AI, and for physician autonomy at the point of care.

(JMIR AI 2026;5:e95742) doi: 10.2196/95742

KEYWORDS

artificial intelligence; clinical decision support; essentially contested concepts; explainability; explainable AI; health AI; interpretability; philosophy of medicine; physician autonomy; transparency

Introduction

The terms transparency, explainability, and interpretability are by now ubiquitous in the health AI literature. They are often used imprecisely. In a 2025 systematic review of 89 papers using the terms, fewer than 20% defined them in any substantive way, and many used explainable and interpretable interchangeably [1]. I suspect that authors who treat the words as synonymous in passing would accept that they do not, in fact, mean the same thing. It is that reflective agreement—that the 3 terms mark genuinely different ideas—that this paper takes as its point of departure.

Some definitions are in order at the outset. By *health AI* I mean machine learning systems intended to inform clinical decisions. By the *kind of claim* a term makes, I mean what a speaker commits themselves to in applying it—whether to a fact about a system's structure, a relation between a system and an observer, or a capacity of an observer. The argument of the paper is that transparency, explainability, and interpretability, though routinely treated as near-synonyms, make claims of these 3 different kinds, and that much of the confusion in the literature follows from running them together.

One caveat about the labels I will use—structural, relational, and functional—should be noted now. They name what settles each claim, its primary criterion of evaluation, not 3 mutually exclusive properties. All 3 concepts involve human observers somewhere: transparency's inspectors, explainability's audience, and interpretability's bounded agent. What distinguishes the

XSL•FO
RenderX

claim types is what the observer must contribute for the claim to be true, a point developed as each is introduced.

It is worth asking why definitions matter at all, given that any we reach will be contingent and revisable. I offer 3 reasons.

First, the terms are relatively novel and, as yet, lack an agreed reference class. When we call a neural network recursive, the word does useful work because it locates the network within a known class of algorithms that share definite properties; recursive has a settled referent. Transparent, explainable, and interpretable do not yet. When we say a health AI system is explainable, it is often unclear whether we are placing it within a class of relevantly similar systems or making a free-floating assertion. The contrast with recursive is precisely the point: the difficulty is not novelty as such—recursive networks are novel too—but the continued absence of a shared referent for these particular terms.

Second, the 3 terms are normative, though not in the sense of being irreducibly moral. They function less like factual predicates than like higher-order principles that tell us which facts about a system are worth attending to, and when—closer to what Cohen [2] called fact-insensitive principles, whose justification does not itself rest on the facts they help us weigh. According to Cohen, when we say a fact *F* confers support on a principle *P*, we are relying, tacitly or not, on another principle that justifies why that fact represents a reason to endorse *P*. That kind of "explaining principle" is a fact-insensitive principle because it holds true regardless of *F*.

Third, health AI operates in a heavily regulated space, and the definitions of these terms increasingly carry legal and operational weight. Regulatory instruments such as the European Union's Artificial Intelligence Act attach obligations to notions of transparency and explanation, which gives their definitions direct consequences for what a compliant system must do [3]. Scientists and clinicians therefore have every reason to care about how these terms are defined, whatever their present appetite for definitional work, because the definitions will be written either way. If those of us who best understand the technology do not set the terms of our own deliberations, others will set them for us, and we will have to live with the result.

I doubt we will ever converge on a single, enduring, all-purpose definition of any of the 3 terms. Like many normative concepts, they appear to be essentially contested in the sense Gallie [4] gave that phrase: concepts whose correct application is permanently and legitimately disputed, not because we lack information but because the disagreement runs internal to the concept itself. This does not foreclose useful agreement: parties who will never share a definition of a contested term can still agree on what kind of claim the term makes, and it is that thinner, sturdier agreement this paper pursues.

With this in view, I can state the paper's aim. I do not try to define transparency, explainability, or interpretability. I ask instead what kind of claim we make when we apply each term. This is a meta-level, second-order question, one step removed from the first-order question of whether a given system is in fact transparent, explainable, or interpretable, which I set aside here. Answering the second-order question first is worthwhile because the 3 terms, I will argue, make 3 different kinds of claims, and much of the confusion in the literature comes from conflating them.

This meta-level framing also distinguishes the paper from existing work. A substantial literature already argues about how to define and evaluate interpretability and explainability, and about whether post hoc explanation is the right goal at all [5-8]. My aim is not to add another definition to that pile but to clarify what sort of thing each definition is trying to be—a structural fact, a relation, or a capacity—so that these debates can be seen to be, in part, disagreements about different categories of claim.

To keep the 3 claim types concrete, I use a single running example from emergency medicine: a machine learning model that predicts, from data routinely gathered in the first hours of an emergency department visit, whether a patient will deteriorate and require escalation to critical care. Call it the deterioration model. It is the kind of system now widely deployed as clinical decision support, and it lets us ask, of one and the same artifact, what it would take for it to be transparent, explainable, and interpretable.

Transparency Claims Are Structural

Overview

To call a system transparent is to make a claim about its accessibility for inspection: roughly, that the box is open. Theories of transparency disagree mainly about how much access is required.

One qualification matters and is easily missed. A model whose code sits on a server no one may examine is, functionally, as closed as one with no code at all, so transparency is indexed to a class of potential inspectors: the open box is open *to someone* positioned to look inside. This indexing can make transparency sound relational in the way explainability will turn out to be, so the difference is worth marking at once. A transparency claim is settled entirely by facts about the system and the access regime surrounding it—what exists to be seen, and who is permitted to see it. No fact about any observer's comprehension enters into its truth conditions: the inspectors need only be positioned at the window, not enlightened by the view. That is the sense in which the claim remains structural.

Here is one very thin and hypothetical version: a model is transparent if a human can directly inspect its parameters, architecture, and decision logic. The visibility boundary is drawn at the model's interior—weights, nodes, and learned structure—and the operative verb is inspect. Whether the observer can make sense of what they inspect is left entirely open. A model with 10 billion weights satisfies this definition: its structure is visible even though it is humanly incomprehensible.

Here is a more demanding and hypothetical one: a model is transparent only if the entire computational process—preprocessing, feature selection, training, and inference—is documented and available for review. This version requires a complete and accessible audit trail, though it too is indifferent to whether the documented choices were sound.

XSL•FO
RenderX

Interestingly, though, not all accounts require us to see the parts of a model. A third and more inventive notion can be drawn from the work of Wachter and colleagues on counterfactual explanation [9]. Their proposal is expressly framed as explanation without opening the black box—the point of a counterfactual is that it does not disclose the model's internals—and they do not themselves offer it as a theory of transparency. But the approach implies a minimal, input-output notion of access worth naming: call it counterfactual transparency, on which a system is transparent to the extent that an observer can determine what went into it and what came out of it. This is interesting because, unlike the previous 2 conceptions, this one allows a fully opaque black box to count as transparent.

Applied to the deterioration model, transparency in the thin sense means its weights and architecture can be inspected; in the fuller sense, that its training data, feature-processing, and validation are documented and auditable; in the counterfactual sense, only that a clinician can see which inputs produced which risk score. None of these guarantees that anyone understands why the score is what it is—that is the province of the next claim. What the 3 conceptions share is the view that disputes about transparency turn on visibility and accessibility, on what can be seen and by whom, not on what the seeing yields.

A model that meets it: A patient I admitted overnight arrests on the ward, and at the morbidity-and-mortality meeting someone asks why the early-warning tool never flagged her. I pull up the model's documentation: the training cohort and its inclusion criteria, the features it ingests, the preprocessing, the validation performance by subgroup, and the version running on our instance. I can see that the model was trained on a population that underrepresented her presentation, and that her chronically low baseline oxygen saturation sat just under the alerting threshold. I could not have computed the output myself, and the vendor's gradient-boosted ensemble means little to me as an object—but nothing was hidden. Everything that produced the number was there to be examined. That is all transparency promises: the box was open, even if what was inside was not, on its own, illuminating.

A model that does not: Our sepsis alert fires on a well-looking patient, and I want to know why. The tool is licensed from a vendor who treats the model as a trade secret. I cannot see the training data, the features, or the thresholds; support tells me only that the algorithm is proprietary and continuously updated. When I ask which of my documentation changes tripped the alert, no one—not the vendor, not our informatics team—can look inside to answer. The system may be excellent, but I have no way to inspect what it did, so I can neither audit it after a bad outcome nor properly defend my reliance on it. The box is welded shut.

Objection: Transparency Claims Are Institutional, Not Structural

The chief objection to a structural reading is that it ignores the moral and political freight the term carries. On this view—associated with work in critical algorithm studies and algorithmic accountability, and articulated in Ananny and Crawford's [10] critique of the transparency ideal—a transparency claim is not a physical description but an institutional and normative one. To say a system is transparent, the objection runs, is rarely just to say the box is open; it is to claim that those responsible for the system have submitted it to oversight, audit, and accountability. Transparency, so understood, is a governance relation, not a structural fact.

Defense: The Structural Core of Transparency Claims

I grant much of this. We do demand transparency for institutional reasons—to build trust, enable audits, and secure accountability—and Ananny and Crawford [10] are right that a naïve equation of transparency with mere disclosure can obscure how little disclosure sometimes accomplishes. But conceding the institutional motivation does not collapse the structural claim into a normative one. My claim is narrower: whatever institutional purposes transparency serves, the mechanism by which those purposes are met is structural. Oversight and audit operate on something—the parameters, the pipeline, the input-output behavior—and it is the accessibility of that something that a transparency claim, at its core, asserts. If we redefine transparency as an institutional posture of openness, we lose the vocabulary to describe the engineering reality on which the posture depends. One can hold very different views about the moral demands of transparency while sharing an account of what, structurally, is being made accessible. The institutional and the structural are not rivals; they answer different questions, and only the second is distinctive of transparency as opposed to accountability in general.

Explainability Claims Are Relational

Overview

To call a system explainable is to make a claim about epistemic mediation: that understanding is successfully conveyed from the system to a particular observer, not merely that some artifact bearing on the system's behavior exists. The analogy is that the box comes with a manual—and, crucially, a manual is written for a reader. Explainability is in this sense irreducibly relational. Unlike transparency, whose truth is settled by the system and the access regime around it without regard to what inspection yields, an explainability claim is incomplete until we specify the observer: their expertise, background knowledge, and inferential capacities. The deterioration model might be explainable to an intensivist familiar with its concepts and wholly unexplainable to the patient whose care it informs, without anything about the model changing. The same gradient runs within the professions. Suppose the model justifies a flag partly by reference to a falling central venous oxygen saturation against a rising vasopressor dose. To an intensivist, that account slots into an existing schema of oxygen delivery and consumption; understanding transfers at a glance. To a generalist covering the ward overnight—no less able a clinician, but not fluent in that vocabulary—the same faithful account may convey nothing actionable; and to the patient it is opaque altogether. One artifact, 3 audiences, 3 different truth values for the explainability claim. That observer-dependence is not a defect of the concept but its defining feature.

XSL•FO
RenderX

A clarification is needed here, because it marks a real limit on the claim. In the philosophy of science, an explanation is not generally treated as relational: on the classical covering-law account and in much of the tradition that followed it, whether some fact explains another is taken to hold independently of any audience [11]. My claim is not about explanation in that sense but about explainability—a system's capacity to produce understanding in a specified class of users. The two come apart. A model can generate a perfectly accurate account of how it reached an output (an explanation) that conveys nothing to the clinician who must act on it (no explainability). To be explicit about what this entails: generating a faithful account of how the output was produced is necessary but not sufficient for explainability. The account must further convey understanding of that process to the specified observer; explainability is achieved at the point of uptake, not at the point of generation.

There is, moreover, a nearby concept, justifiability, which Muralidharan and colleagues [12,13] develop on an explicitly Rawlsian, relational model of interpersonal justification. The 2 notions answer different questions. Explainability, as the term is actually used in health AI, concerns whether a specified observer comes to understand how the system reached its output. Justifiability concerns whether the decision itself can be shown to be warranted—whether the patient can be given adequate reason to regard it as correct or, where it is wrong, to see why it is wrong. Justifiability is in that sense explainability's normative cousin: what the patient is owed, on Muralidharan and colleagues' [12,13] account, is not access to the model's internals but a justification they could reasonably accept. The taxonomy here concerns the former; the latter marks the direction in which an account of what patients are owed would have to develop.

Theories of explainability concern what makes a valid bridge between machine output and human cognition. Three families are instructive, and it is worth being explicit about the sense in which each is relational—including where the relation can fail.

Post hoc additive explanation (PAE) methods attach feature-attribution accounts to a black box model after the fact. Local interpretable model-agnostic explanations (LIME) fits a simple surrogate model that is faithful to the black box in the neighborhood of a single case [14]; Shapley additive explanations (SHAP) allocates a prediction among the input features using Shapley values drawn from cooperative game theory [15]. The machinery differs—local surrogate fitting in one, an axiomatically grounded attribution scheme in the other—but both deliver the same kind of artifact: a ranked list of each feature's contribution to a given output. Applied to the deterioration model, either method would tell a clinician that this patient's risk score was driven mainly by a rising respiratory rate, a falling oxygen saturation, and an elevated lactate. PAE's relationality is real but conditional: the method presupposes an observer for whom the reported features are meaningful, and offers no guarantee that they will be. If the salient feature is an opaque embedding dimension rather than lactate, the bridge is built but lands nowhere. PAE is relational in its ambition—it aims to convey contributions to a user—but it does not ensure arrival.

Counterfactual explanation (CE) identifies the smallest change to the input that would have altered the output [9]. For the deterioration model, a CE answers: had this patient's oxygen saturation been 3 points higher, the model would not have flagged them. The bridge runs from the actual decision to an actionable alternative, answering what would have to change rather than why. Like PAE, its relationality is genuine but incomplete: the counterfactual illuminates only if the altered feature is one the observer can recognize and act on. A counterfactual stated over an uninterpretable feature is as relational in form, and as empty in effect, as an unreadable saliency map.

Concept bottleneck models (CBMs) take a stronger, architectural approach: the model is forced to route its prediction through a layer of human-specified concepts before producing an output [16]. The deterioration model, built as a CBM, would be required to express itself in a fixed clinical vocabulary—signs of respiratory distress, hemodynamic instability, and evidence of infection—before returning a risk score, and could not appeal to latent features that lack a name in that vocabulary. Here the relation to the observer is secured architecturally rather than merely attempted: the model must speak in terms the observer already possesses. Even this guarantee needs qualification. What the architecture ensures is that the prediction routes through the concept vocabulary; it does not ensure that the learned concept representations carry only the clinical meaning their labels advertise, and learned concepts can encode additional information that leaks past those labels. With that caveat in mind, this is the clearest case of the relational thesis, and also where its cost is most visible. The bridge is only as wide as the concept set, which is fixed before the model encounters any case. Confronted with a presentation whose relevant features fall outside that set—a novel toxidrome or an atypical presentation the vocabulary was never given words for—a CBM cannot represent what it is missing. The failure is not that the modeler forgot to ask some contextual question; it is that the concept vocabulary simply lacks the terms, and a model that can speak only in fixed concepts cannot flag the presence of something it has no word for.

These 3 families differ sharply in method, but each treats explainability as something that emerges in the interaction between system and observer, and each locates its central dispute in how to make that interaction succeed. It would overstate the case to say that all 3 guarantee understanding: PAE and CE plainly do not, and even a CBM's guarantee is only as good as its concept set. The unifying point is weaker but still substantive. All 3 treat explainability as audience-indexed—as a claim about a bridge to a specified observer rather than a property of the artifact alone. What separates them is how much of the burden of building that bridge they place on the model's architecture as opposed to a post hoc procedure.

A product that meets it: A deterioration tool flags a patient in the resuscitation bay as high-risk. Instead of a bare score, it shows its reasoning in terms I already hold: a respiratory rate climbing across the last 3 sets of observations, a saturation falling despite increasing oxygen, and a lactate of 4.1. That is a language I think in. I look at the patient, agree the trajectory is worrying, and escalate. The explanation did its work because

XSL•FO
RenderX

it was pitched to the person who has to act on it; understanding actually transferred.

A product that does not: A different tool flags the same kind of patient, and its vendor has clearly tried: the interface is polished, the feature names are in plain clinical English, and it presents a tidy ranked list of contributors. But when I press on how the score is actually reached, the honest answer is that it turns on a high-order interaction among several hundred variables—a distributed pattern in the data with no counterpart in how I reason about a patient. The clean clinical labels on the dashboard are a post hoc gloss laid over that pattern; they read well but do not track what the model did. So, I am left choosing between an account that is faithful and unintelligible and one that is intelligible and confabulated. This is not something better design can fix, because there is no clinical concept on my side of the bridge for a faithful explanation to land on. The artifact is not merely badly presented; it is unexplainable to me in principle.

Objection: Explainability Claims Are Formal, Not Relational

The main objection to a relational reading is that it makes explainability subjective, tethering it to human psychology rather than to a property of the system. On this view, an explainability claim is formal or generative rather than relational: formal, in that its truth depends only on features of the explanatory artifact itself, above all its mathematical fidelity to the underlying model; and generative, in that a system counts as explainable if it can generate such an artifact, whether or not any particular observer grasps it. Whether the clinician happens to possess the background to understand a saliency map, the objection runs, is irrelevant to whether the explainability claim is true. This is the sense in which industry practice often calls a method explainable when it faithfully reports how the model reached its output, full stop.

Defense: Why the Relational Reading Better Describes the Practice

My defense is not that the formal reading is incoherent but that it misdescribes what these methods are for. Note first what the objection gets right: if a purported explanation does not even faithfully track how the model reached its output—if it is systematically inaccurate—we would withhold the word explainable regardless of any observer. Fidelity is necessary. The question is whether it is sufficient, and here the relational reading does the better descriptive work. Each of the 3 families is built to deliver something to somebody: contributions a user can weigh, alternatives a user can act on, and concepts a user already holds. Strip out the user, and one cannot say what any of them is optimizing. A mathematically perfect saliency map that no clinician can read is, in the terms drawn earlier in this section, a successful explanation but a failed instance of explainability. The formal account has no way to register that second failure; the relational account names it exactly. The claim, then, is not merely that relationality is desirable but that it is descriptively indispensable: it is the only reading on which the disputes between these methods—which are disputes about how best to reach an observer—are intelligible as disagreements about the same thing [8]. It is also the reading toward which the human-centered strand of the explainable AI literature has independently converged: work in that tradition treats explanation as something designed for, and evaluated against, the needs of particular users in particular contexts, rather than as a free-standing property of the artifact [17,18].

A final practical dimension of the relational reading deserves notice. In clinical use, understanding is not an end state but an input to action. The point of conveying how the model reached its output is to put the clinician in a position to interrogate it: to check the cited features against the patient in front of them, to notice when the account does not fit, and to decide on that basis whether to follow, verify, or override the tool. An explainability claim should accordingly be evaluated not only on whether understanding transfers but on whether the resulting understanding supports this kind of critical engagement—a disposition that recent empirical work has begun to measure directly, separating users' verification behavior, motivation, and reflection when acting on AI outputs [19]. An explanation that produces fluent deference rather than checkable understanding fails its observer as surely as an unreadable one, and more insidiously [8].

Interpretability Claims Are Functional

Overview

To call a system interpretable is to make a claim about simulability: that a bounded human agent can, in principle, follow the system's procedure step by step and rederive its output by the same path. The analogy is that one can take the mechanism apart and work the gears oneself. This needs immediate qualification, because a naïve version invites an obvious counterexample. Consider a pocket calculator. Its algorithm for the natural logarithm is about as simple and well-understood as algorithms come, yet no one can compute ln(5) in their head. If interpretability required that a human actually carry out the whole computation mentally, the calculator would come out uninterpretable, which is absurd. The lesson is that simulability is not a claim about mental horsepower or working memory. It is a claim about the surveyability of the procedure: each step is the kind of operation a human can follow and check, and the sequence of steps is short and structured enough that a human could, in principle and with pen and paper, walk the same path. Interpretability, so understood, is a property of the procedure's form, not of any observer's capacity to hold the whole computation in mind at once.

Two consequences of this formulation should be made explicit, because the phrase "in principle" otherwise promises more than it can keep. Any finite computation—the forward pass of a deep network included—could be traced by hand in principle, given unlimited time and paper. What distinguishes procedures is how far they exceed what a bounded agent can actually survey, and this means that interpretability on the present account is scalar rather than binary. A linear model with 5 coefficients sits near one end of the scale. A linear model with a million coefficients has exactly the same form yet sits far down the scale, because surveyability degrades with size even when structure is held constant; tracing it is possible only in the way that counting the grains in a sandbag is possible. A deep network sits at the far extreme on both counts, its procedure vast and unstructured

XSL•FO
RenderX

alike. Interpretability claims are therefore claims of degree, and the threshold that matters—surveyable enough for the bedside, for an audit, or for a regulatory submission—is set by the task and its time frame, not by the mathematics.

So construed, interpretability draws a clean line between explainability and itself. Explainability asks whether a human can understand what the model is doing; interpretability asks whether a human could, in principle, do what the model is doing—reproduce the output by following the same procedure. A sparse linear model with 5 coefficients is interpretable because a clinician could, in principle, perform the weighted sum by hand [20]; a deep network with millions of parameters is not, however many post hoc explanations we attach, because there is no surveyable procedure a human could follow [6]. A CBM may be highly explainable—its intermediate concepts make sense—while failing to be interpretable, if the mapping from concepts to output is itself an unsurveyable learned function [16].

Theories of interpretability disagree about where human cognitive limits fall and which model families stay within them.

Sparse linear models such as the least absolute shrinkage and selection operator (LASSO) restrict the model to a few nonzero coefficients, so that the output is a weighted sum an observer can check by hand [20]. This is the archetype of the bedside risk score: the deterioration model, rendered as a handful of weighted findings, is not merely described by a rule the clinician can compute—it is that rule. The locus of dispute is expressiveness: linear combinations cannot capture the nonlinear interactions endemic to biology. The model is simulable precisely because it is austere. It is worth being honest that a weighted sum is not a full procedure in the way a clinical protocol is—it has no ordering, no steps that must precede others—so what such a model makes interpretable is the rule's structure and proportions, not a sequence of operations. That is usually what is wanted of a score, but it is a narrower notion of interpretability than the word carries in some other settings.

Shallow decision trees encode logic as a branching sequence of binary splits, and the observer traces a single path from root to leaf [21]. The locus of dispute is depth: cognitive tractability collapses as trees grow. Beyond a handful of splits, the branching structure exceeds working memory, and the tree becomes something a human can no longer survey in one pass. A shallow tree over the deterioration model's inputs—is the respiratory rate above a threshold, and then is the lactate above a threshold—is walkable; a deep one is not.

Generalized additive models decompose the prediction into a sum of smooth, single-feature shape functions, 1 per input, each of which can be plotted and mentally aggregated [22]. The locus of dispute is interactions: strict additivity forbids learned interactions, so any interaction must be specified by hand. The architecture buys simulability at the price of automated discovery. For the deterioration model, a generalized additive model would let a clinician read off how risk rises with respiratory rate and, separately, with lactate; but if what actually matters is the joint effect of a rising lactate in a patient who is also hypotensive—an interaction no single feature captures—the additive model cannot find it unless the modeler already suspected it and built it in.

Rashomon sets are best introduced not as a fourth interpretable model family but as a complementary framework [7,23]. Rather than commit to one model, a Rashomon analysis characterizes the whole set of models that perform near-equivalently on a task—a set that can contain sparse linear models, shallow trees, generalized additive models, deep networks, and anything else that clears the performance bar, at every level of interpretability. That such sets are often populated by strikingly diverse architectures raises a question the individual families do not: if many different models predict deterioration equally well, which features are genuinely essential and which are one model's incidental preference? What a Rashomon analysis yields is therefore less an artifact of any single model than an interpretation of what the adequate models share—and, for present purposes, a principled way of asking whether an interpretable member of the set can be chosen without sacrificing performance [7].

There is a further point worth drawing out, because it closes a loop opened earlier. The Rashomon finding—that many nonequivalent models fit the data about equally well—is a kind of empirical echo of Gallie's [4] thesis in the technical domain. Just as an essentially contested concept admits multiple defensible applications that no amount of further information adjudicates, a Rashomon set admits multiple defensible models that the data alone does not adjudicate between. The analogy should not be pressed too hard. Gallie's [4] contests are normative disputes among the users of a concept, sustained by differing values; a Rashomon set is a mathematical fact about a hypothesis space, and its members can in principle be separated by further data, by shifted populations, or by criteria other than accuracy. The echo is structural rather than exact: in both cases an underdetermined choice remains after the evidence has said all it can, and something other than the evidence must settle it.

Across the 3 model families, interpretability turns on the same requirement: that the observer be able, in principle, to rederive the output by following the procedure. The Rashomon framework does not supply such a procedure itself; its contribution is to show how many procedures, surveyable and unsurveyable alike, the data equally supports. That common requirement—rederivability by a bounded agent—is what makes interpretability functional.

A product that meets it: For chest pain, I use a scoring rule with a handful of weighted criteria. When it returns low risk, I can add the points up on the edge of the chart myself and see exactly how it got there. That matters in the case where the inputs mislead—the patient whose story worries me more than the score does. Because I can reproduce the calculation, I can see which input is carrying the result, judge that it underweights this presentation, and override it on my own authority. The rule does not ask for my trust; it invites my checking.

A product that does not: Our imaging triage model returns a probability of critical deterioration from a deep network with millions of parameters. Even with attributions bolted on, there is no procedure I can walk: I cannot, in principle or with pen

XSL•FO
RenderX

and paper, trace the computation from input to output and arrive at the same number. When the model and I disagree, I am not weighing my reasoning against a procedure I can survey—I am deciding whether to defer to an oracle. It may be well-calibrated and well-explained, but it is not something I can do myself, and so it cannot support my judgment the way the score does; it can only stand in for it.

Objection: Interpretability Claims Are Semantic, Not Functional

One might object that interpretability is not a functional claim about rederivation at all, but a semantic or ontological one about representation—and these are worth separating, since the objection has a weaker and a stronger form. The semantic version says that what makes a model interpretable is that its internal variables mean something to the observer: the features refer to things the observer can recognize. The ontological version is stronger, holding that interpretability requires the model's structure to correspond to the actual structure of the phenomenon—that the model carve the world at something like its real joints, not merely traffic in familiar-seeming labels. On either version, the objection continues: a merely functional account is inadequate: a procedure can be perfectly surveyable and still trade in variables that are meaningless (the semantic worry) or that misrepresent the underlying biology (the ontological worry). And even where the representation seems to make sense, that appearance may be an artifact of the observer's cognitive biases rather than a fact about the model. A clinician who can trace every split of a decision tree whose nodes are spurious correlations has simulated a procedure without understanding anything real.

Defense: The Functional Distinction of Interpretability Claims

The objection is appealing, but it ultimately folds interpretability back into explainability. Semantic alignment—the model's variables meaning something to the observer—is just a strong, built-in explainability claim: it makes the model understandable to a cognitive agent, which is precisely what the previous section called relational. If interpretability is to remain a distinct category rather than a special case of explainability, it must demand something the semantic account does not. What it demands is that the observer be able to complete the inference themselves. I earlier reached for the term closure for this, and it is worth stating plainly what is and is not meant. I do not mean deductive closure—working out everything entailed by what one already knows. I mean something more modest: that the observer can independently carry the inference from inputs to output through to its end, without having to take the model's word for any step. Semantic alignment still asks the human to trust the model's map of the world; functional simulability lets the human redraw the map and check that it leads to the same place. The ontological worry—that the variables might misrepresent the biology—is real, but it is a worry about whether the model is any good, not about whether it is interpretable. An interpretable model can be interpretably wrong, and its being surveyable is exactly what lets us see that it is wrong. This is why interpretability figures so prominently wherever the question is how far to trust a system at all, including much of the regulatory debate.

Discussion and Conclusion

This paper has done one limited thing. It has not defined transparency, explainability, or interpretability; it has argued that the 3 terms make 3 different kinds of claims—structural, relational, and functional, respectively—and that keeping them apart clarifies a literature that routinely runs them together. Table 1 summarizes the taxonomy, including the principal accounts or families canvassed for each term; as noted in the Introduction, the claim-type labels identify each term's primary criterion of evaluation rather than mutually exclusive properties.

If the taxonomy is right, several practical consequences follow for how health AI is built, evaluated, reported, and regulated. Because the 3 terms make different kinds of claims, they call for different kinds of evidence, and conflating them produces category errors in evaluation. A transparency claim, being structural, is discharged by disclosure and audit: one shows the parameters, the pipeline, or the input-output surface. It is not answered by a user study, and a system can be fully transparent while remaining unexplained and uninterpretable. An explainability claim, being relational, cannot be established by exhibiting an artifact alone; because it is indexed to an audience, its evaluation requires specifying that audience and testing whether understanding actually transfers—a reporting standard that names the intended user and the task, not merely the method. An interpretability claim, being functional, is settled by the model's form: whether a bounded agent could rederive the output by a surveyable procedure—a judgment of degree, calibrated to the task. These distinctions bear directly on regulation. A requirement that a device be explainable, if it is to mean anything, must specify explainable to whom; a requirement of transparency can be met by documentation that leaves clinicians none the wiser; and a requirement of interpretability constrains the model class itself. Regulators who use the terms interchangeably will write requirements that cannot be jointly satisfied, or that are satisfied vacuously.

Existing governance instruments illustrate both the need for these distinctions and their beginnings. The World Health Organization's guidance on the ethics and governance of AI for health gathers transparency, explainability, and intelligibility into a single guiding principle, leaving unspecified precisely the distinctions among them that this paper urges [24]. The joint guiding principles on transparency for machine learning–enabled medical devices issued by the US Food and Drug Administration, Health Canada, and the United Kingdom's Medicines and Healthcare Products Regulatory Agency define transparency in expressly audience-indexed terms, as the clear communication of appropriate device information to relevant audiences [25]. In the present taxonomy, that definition welds a structural claim to a relational one, and implementing it would be easier if the 2 were named separately: what must be disclosed is one requirement; what a named audience must come to understand is another. One disclosure of my own belongs here rather than only in the declarations below: I am developing a commercial clinical decision support platform, so

XSL•FO
RenderX

recommendations about how such systems should be evaluated and regulated concern a class of products in which I hold an interest, and the reader should weigh them accordingly.

The distinctions also bear on a question of growing importance: the effect of health AI on physician autonomy. A recent scoping review finds the relationship understudied and its mechanisms contested [26], and the ethics literature increasingly frames the issue as one of preserving clinicians' capacity for independent judgment under algorithmic influence [27,28]. The taxonomy suggests that the 3 properties support autonomy in different ways, and that which one matters most is context-dependent. In the emergency department, interpretability may matter most: a scoring rule a physician can compute and check preserves the capacity for independent judgment in a way a black box prediction, however well explained, does not, because the physician can rederive it and, where warranted, override it. In radiology, explainability may take priority: concept-based explanations that align with diagnostic ontology can enter the existing reasoning workflow rather than displace it. In regulatory submission, transparency may be the threshold: access to training data, validation protocols, and architectural documentation is what makes any subsequent scrutiny possible. Autonomy is not served by maximizing all 3 at once, which is often impossible, but by identifying which claim the clinical context actually requires.

I have deliberately stopped short of saying which definitions we should adopt. My own view, grounded in day-to-day emergency practice, is that health AI demands practicality above universality: I lean away from all-purpose definitions—Gallie [4] was right that some concepts resist that kind of closure [4]. But the argument of this paper does not depend on that preference. Whatever definitions we ultimately reach, it helps to know first what kind of thing we are trying to define. Getting the category right is the modest but necessary condition of getting anything else right.

Table 1. The 3 claim types compared.

	Transparency	Explainability	Interpretability
Claim type	Structural	Relational	Functional
Core question	What can be inspected about the system?	Is understanding conveyed to a specified observer?	Can a bounded human rederive the output by following the procedure?
Principal accounts or families	Parameter-level inspection; full process documentation and audit; counterfactual (input-output) access	Post hoc additive explanation (LIME[a] and SHAP[b]); counterfactual explanation; concept bottleneck models	Sparse linear models (eg, the LASSO[c]); shallow decision trees; generalized additive models; Rashomon sets as a complementary framework
Health AI example	Deterioration model: weights, training pipeline, and validation are available for audit	Deterioration model: a concept- or feature-based account a clinician can actually use	Deterioration model expressed as a bedside score computable by hand
Evaluation criterion	Disclosure and audit of the system's internals or input-output surface	Whether understanding transfers to the named audience for the task	In-principle surveyability of the procedure by a bounded agent; a matter of degree
Main limitation	Visibility does not entail understanding; a fully visible model can be incomprehensible	Audience-indexed and not guaranteed by any artifact; fidelity is necessary but not sufficient	Buys surveyability by restricting model expressiveness; scalar, and narrower than some common usage

[a]LIME: local interpretable model-agnostic explanations.

[b]SHAP: Shapley additive explanations.

[c]LASSO: least absolute shrinkage and selection operator.

Acknowledgments

The author would like to thank the scientific editor and the 3 anonymous reviewers for their careful and generous engagement with the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

Data availability is not applicable to this article, as no new data were generated or analyzed. This is a conceptual viewpoint and does not rely on an underlying dataset.

XSL•FO
RenderX

Generative AI Use

Generative AI (Claude; Anthropic) was used to assist with bibliographic research and copyediting. All substantive arguments, claims, and final wording are the author's own, and the author reviewed and takes full responsibility for the content of the manuscript.

Authors' Contributions

TR is the sole author and was responsible for conceptualization, writing—original draft, and writing—review and editing.

Conflicts of Interest

TR is developing Vetta, a commercial AI-enhanced clinical decision support platform, in which the author holds a financial interest. Vetta is not named, described, or evaluated in this manuscript, and no Vetta-related data or materials were used. Because the manuscript makes recommendations bearing on the evaluation and regulation of clinical decision support systems as a class, this interest is also acknowledged in the Discussion. The author declares no other conflicts of interest.

References

1. Noor AA, Manzoor A, Mazhar Qureshi MD, Qureshi MA, Rashwan W. Unveiling explainable AI in healthcare: current trends, challenges, and future directions. WIREs Data Min Knowl Discov. 2025;15(2):e70018. [doi: 10.1002/widm.70018]
2. Cohen GA. Facts and principles. Philos Public Aff. 2003;31(3):211-245. [doi: 10.1111/j.1088-4963.2003.00211.x]
3. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union. 2024;1689:1-144. [doi: 10.2139/ssrn.4724557]
4. Gallie WB. Essentially contested concepts. Proc Aristot Soc. 1956;56(1):167-198. [doi: 10.1093/aristotelian/56.1.167]
5. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. Preprint posted online on February 28, 2017. [doi: 10.48550/arXiv.1702.08608]
6. Lipton ZC. The mythos of model interpretability. Commun ACM. 2018;61(10):36-43. [doi: 10.1145/3233231]
7. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-215. [FREE Full text] [doi: 10.1038/s42256-019-0048-x] [Medline: 35603010]
8. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-e750. [FREE Full text] [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]
9. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv J Law Technol. 2018;31(2):841-887. [FREE Full text]
10. Ananny M, Crawford K. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society. 2016;20(3):973-989. [doi: 10.1177/1461444816676645]
11. Hempel CG. Aspects of Scientific Explanation and Other Essays in the Philosophy of Science. New York, NY. Free Press; 1965.
12. Muralidharan A, Savulescu J, Schaefer GO. AI and the need for justification (to the patient). Ethics Inf Technol. 2024;26(1):16. [FREE Full text] [doi: 10.1007/s10676-024-09754-w] [Medline: 38450175]
13. Muralidharan A, Savulescu J. All you need is… justification: algorithmic justifiability trumps transparency. Ethics Inf Technol. 2026;28(2):18. [doi: 10.1007/s10676-026-09892-3]
14. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. 2016. Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 August 13-17:1135-1144; San Francisco California USA. [doi: 10.1145/2939672.2939778]
15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017. Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 December 4-9:4768-4777; Long Beach California USA.
16. Koh PW, Nguyen T, Tang YS, Mussmann S, Pierson E, Kim B, et al. Concept bottleneck models. 2020. Presented at: Proceedings of the 37th International Conference on Machine Learning; 2020, July 13-18:5338-5348; Virtual.
17. Liao QV, Varshney KR. Human-centered explainable AI (XAI): from algorithms to user experiences. arXiv. Preprint posted online on October 20, 2021. [doi: 10.48550/arXiv.2110.10790]
18. Ehsan U, Riedl MO. Human-centered explainable AI: towards a reflective sociotechnical approach. In: HCI International 2020 ' Late Breaking Papers: Multimodality and Intelligence. Lecture Notes in Computer Science. Cham. Springer; 2020:449-466.
19. Lau GR, Low WY, Tay L, Guevarra YA, Gašević D, Hartanto A. Understanding critical thinking in generative artificial intelligence use: development, validation, and correlates of the Critical Thinking in AI Use Scale. Comput Hum Behav Rep. 2026;22:101103. [doi: 10.1016/j.chbr.2026.101103]
20. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol. 1996;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]

XSL•FO
RenderX

21. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Belmont, CA. Wadsworth; 1984.
22. Hastie TJ, Tibshirani RJ. Generalized Additive Models. London. Chapman & Hall; 1990.
23. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16(3):199-215. [doi: 10.1214/ss/1009213726]
24. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva. World Health Organization; 2021.
25. US Food and Drug Administration, Health Canada, Medicines and Healthcare Products Regulatory Agency. Transparency for machine learning-enabled medical devices: guiding principles. US Food and Drug Administration. URL: https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles [accessed 2026-09-01]
26. Grosser J, Düvel J, Hasemann L, Schneider E, Greiner W. Studying the potential effects of artificial intelligence on physician autonomy: scoping review. JMIR AI. 2025;4:e59295. [FREE Full text] [doi: 10.2196/59295] [Medline: 40080059]
27. Funer F, Wiesing U. Physician's autonomy in the face of AI support: walking the ethical tightrope. Front Med (Lausanne). 2024;11:1324963. [FREE Full text] [doi: 10.3389/fmed.2024.1324963] [Medline: 38606162]
28. Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020;46(3):205-211. [FREE Full text] [doi: 10.1136/medethics-2019-105586] [Medline: 31748206]

Abbreviations

CBM: concept bottleneck model
CE: counterfactual explanation
LASSO: least absolute shrinkage and selection operator
LIME: local interpretable model-agnostic explanations
PAE: post hoc additive explanation
SHAP: Shapley additive explanations

Edited by A Coristine; submitted 19.Mar.2026; peer-reviewed by A Muralidharan, J Grosser, E Karampinis, GR Lau; comments to author 22.May.2026; revised version received 25.Aug.2026; accepted 30.Aug.2026; published 19.Nov.2026

Please cite as:
Ringer T
What Kind of Claims Are Transparency, Explainability, and Interpretability? A Definitional Taxonomy for Health AI
JMIR AI 2026;5:e95742
URL: https://ai.jmir.org/2026/1/e95742
doi: 10.2196/95742
PMID: 42669108

©Tessa Ringer. Originally published in JMIR AI (https://ai.jmir.org), 19.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX