

Contextual and Global Falsification of Scientific Models

An Integrated Theory of Epistemic Validity

Abstract

The classical and widely received Popperian theory of falsification operates with an idealized binary picture of scientific rationality, according to which a theory is in principle undermined by a conflicting observation (Popper [1934] 1959). Even though Popper himself acknowledges methodological side conditions, this binary-eliminative structure remains too coarse for analyzing modern modeling practice. Models such as Newtonian mechanics, classical thermodynamics, or contemporary climate models are empirically limited in certain subdomains yet remain epistemically indispensable in others.

This paper develops an integrative framework of model validity as a philosophy-of-science specialization of model management under finite conditions. It combines insights from Popper, Kuhn, Lakatos, da Costa/French, and contemporary model theory; distinguishes global from contextual falsification; and introduces the epistemic enabling space $E(t)$. $E(t)$ renders explicit the methodological, technical, and institutional conditions under which models can be formulated, tested, stabilized, and replaced by alternatives.

"Falsification" is understood here not as a purely logical criterion of truth or falsity but as a rational-pragmatic mechanism of model evaluation. The decisive issue is not merely whether a model's performance declines in a given domain, but whether available alternatives within the same relevant domain prove stably more epistemically competitive over time. The proposed reconstructive evaluative structure links approximate truth, explanatory power, and model costs without treating scientific rationality as generally algorithmically computable. A case study of climate models illustrates how ensemble methods, parameterizations, and probabilistic weighting procedures contribute to domain refinement. The theory thereby reconstructs central structures of modern modeling practice and explains why models remain stable despite partial falsifications, when they are revised, and under which conditions global model replacement occurs.

Keywords

contextual falsification; global falsification; model validity; model choice; approximate truth; truthlikeness; scientific models; epistemic costs; model revision; epistemic enabling space; philosophy of science; scientific realism; simulation models; climate models; epistemics

Stefan Rapp

Independent Researcher

Last revised: 27 April 2026

<https://orcid.org/0009-0004-0847-9164>

<https://doi.org/10.5281/zenodo.17714966>

Epistemics Project: <https://www.epistemics.de>

© 2026 Stefan Rapp — Licensed under CC BY-NC-ND 4.0

Table of Contents

1. From Binary Falsification to Domain-Relative Model Validity.....	3
2. State of Research.....	5
2.1 Popper: Falsification as a Binary Ideal.....	5
2.2 Duhem-Quine: Underdetermination of Empirical Tests.....	5
2.3 Kuhn: Anomalies without Elimination.....	5
2.4 Lakatos: Research Programs.....	5
2.5 Model Theory.....	6
2.6 Approximate Truth.....	6
2.7 Research Gap.....	6
2.8 Contribution of This Paper.....	7
3. Basic Concepts.....	8
4. Criteria of Modern Model Evaluation within the Proposed Framework.....	13
5. Proposal: Two-Stage Falsification Structure.....	15
5.1 Limiting Routes of Immunization.....	16
6. When Models Disappear.....	17
7. Why Models Persist.....	18
7.1 Robustness through Model Families.....	18
7.2 Contexts of Use.....	18
8. Case Study: Climate Models.....	19
8.1 Climate Models as Ensemble Structures.....	19
8.2 Parameterization as an Epistemic Operation.....	19
8.3 Bayesian Updating and Adaptive Modeling.....	20
8.4 Domain Structure.....	21
8.5 Miniaturized Formal Illustration.....	21
8.6 Consequences for Epistemic Status.....	22
9. Approximate Truth.....	23
10. Why an Integrated Theory Has Been Lacking.....	25
10.1 Historical Reasons.....	25
10.2 Logical Reasons.....	25
10.3 Institutional Reasons.....	25
11. The Epistemic Enabling Space.....	26
11.1 Dynamics of the Epistemic Enabling Space.....	26
11.2 Structure of the Epistemic Enabling Space $E(t)$	27
12. The Decision Logic of Model Choice.....	28
12.1 Structure of the Utility Function.....	28
12.2 Scaling and Empirical Anchoring of the Utility Function.....	30
12.3 Diagnostic Protocol of Contextual and Global Falsification.....	31
13. Relation to Scientific Realism.....	33
14. Falsification as Domain-Relative Revision of Validity.....	34
Local Philosophy-of-Science Glossary for This Paper.....	35
Bibliography.....	38
Appendix: Didactic Examples of the Two-Stage Falsification Structure.....	40

1. From Binary Falsification to Domain-Relative Model Validity

For decades, Popper's theory of falsification ([1934] 1959) shaped both the public and the academic image of scientific rationality. In its classical and widely received form, falsification appears as an idealized binary test of a theory's scientific standing: a theory comes under pressure with respect to its scientific acceptability if one of its predictions empirically fails. The practice of the natural and social sciences, however, presents a more complex picture. Many models are idealized, only partially representational, and empirically limited in subdomains, yet remain epistemically stable. This paper therefore develops an integrated theory of scientific model validity that brings together central insights from Popper, Kuhn, Lakatos, da Costa/French, and contemporary model theory within a common structure of rational model evaluation.

Examples include Newtonian mechanics, which fails at relativistic velocities (Einstein 1905); classical thermodynamics, which is microscopically imprecise; and climate models, which are continuously recalibrated (Oreskes et al. 1994). At the same time, there are models that have disappeared entirely, such as phlogiston theory or the classical ether.

The core problem is this: Under what conditions does falsification lead to the elimination of a model—and when does a model remain epistemically stable despite partial falsification?

In what follows, this problem is treated as a special case of epistemics, that is, as a special case of model management under finite conditions (Rapp 2026a). The paper is thus to be understood as a module within a broader research program in which friction, domain structure, search, and revision are developed in their own dedicated works. The present study transfers this architecture to the case of scientific falsification. Scientific models possess no unlimited validity; they are stabilized, tested, restricted, and revised within particular domains. The general concepts of model, validity, domain, costs, and revision are therefore applied here to the special case of scientific falsification.

In this framework, falsification appears not primarily as a simple operation of elimination, but as a specific form of epistemic friction in the sense of a boundary signal of finite viability (Rapp 2026b): it indicates that the prior stabilization of a model in a given domain is no longer straightforwardly viable. Friction functions here as a procedural indicator of a strained utility structure. It arises when empirical deviations weaken $AT(M, D, t)$, explanatory deficits reduce $EK(M, D, t)$, model costs $C(M, D, t)$ rise, or available alternatives M^* become relatively more competitive. Whether this leads to local revision, restriction of the domain of validity, or global model replacement depends on the model's stable relative competitiveness within $E(t)$.

This paper develops an integrated theoretical framework that reconstructs this question explicitly within a unified model. The approach distinguishes global from contextual falsification, describes the epistemic enabling space $E(t)$, and integrates approximate truth as a graded, domain-specific evaluative quantity. Falsification is understood not as a purely logical operation on truth values but as a rationally reconstructible change in model validity within a real scientific enabling space. The decisive issue is not the isolated decline in the performance of a model, but its relative competitiveness against available alternatives within the same relevant domain. The theory is primarily descriptive and reconstructive: it reconstructs the actual modeling practice of contemporary sciences and is illustrated through climate models.

The utility function $U(M, D, t)$ introduced below is to be understood primarily in reconstructive terms. It does not describe a general computational machinery of scientific rationality but renders explicit how model choice in actual sciences can be reconstructed when approximate truth, explanatory power, and model costs are weighed against one another. Within this paper, $U(M, D, t)$ formalizes the domain-specific validity of scientific models under finite conditions. At the same time, the same structure admits a weakly normative reading, since it brings into view the conditions under which model decisions can count as rationally intelligible. The theory is thus both a descriptive reconstruction of modeling practice and a framework for epistemic optimality, without thereby deriving a universal decision formula.

2. State of Research

2.1 Popper: Falsification as a Binary Ideal

Popper ([1934] 1959) formulates falsification as a methodological counterpart to verification: scientific theories must be such that they can fail against possible observations. In its classical, widely received Popperian structure, falsification accordingly appears as an idealized binary test of a theory's scientific standing. This structure presupposes strong conditions: relatively global claims to validity, unambiguous theory-observation mappings, and sufficiently isolable test conditions. Idealized, approximative, and simulation-based models in particular fit only partially within this binary structure. The tension between Popper's ideal and actual modeling practice has been described many times in the literature (Hacking 1983; Cartwright 1983), but has rarely been developed into a systematic theory of contextual falsification.

2.2 Duhem-Quine: Underdetermination of Empirical Tests

The Duhem-Quine problem shows that empirical tests rarely target isolated theories; they typically target bundles of modeling assumptions, auxiliary hypotheses, measurement practices, and background conditions. The framework developed here adopts this insight but reconstructs it in a domain-relative manner: what matters is not only which assumption is affected by a deviation, but whether this leads to local revision, domain restriction, or global model replacement. Falsification thus appears not as the immediate logical negation of an individual model but as a structured change of model validity within particular conditions of testing and application.

2.3 Kuhn: Anomalies without Elimination

Kuhn (1962) shows that scientific paradigms survive anomalies. He does not, however, provide a finely structured theory of model-relative falsification. The shortcoming consists especially in the fact that Kuhn does not specify how anomalies distribute across subdomains and why they barely affect some models while undermining others. The literature emphasizes Kuhn's historical and sociological perspective, but a formally precise model-theoretic treatment of how subdomains are handled is lacking (cf. Lakatos 1970; Weisberg 2013).

2.4 Lakatos: Research Programs

Lakatos (1970) explains stability through a "hard core" but does not specify when models actually disappear. It also remains unclear how competing programs are to be weighed against one another in detail, especially when both are partially, but not globally, falsified. Here, in particular, it becomes evident that Lakatos's program structure provides no formal instrument for evaluating domain-specific approximate truth.

2.5 Model Theory

Model-theoretic approaches (Cartwright 1983; Weisberg 2013; Morgan and Morrison 1999) emphasize the idealized and domain-specific character of scientific models, but offer no theory of how falsification operates within such structures or how models can be systematically weighed against each other. Theories of explanation complement this picture, showing that explanatory power can be reconstructed in different ways: as the unification of many phenomena under a small set of explanatory patterns (Kitcher 1981, 1989), or as causal-interventionist insight into possible changes (Woodward 2003). Similar insights have also been developed in debates about model pluralism, idealization, robustness, and simulation. Work on pluralistic scientific practice, on isolating modeling strategies, and on simulation-based knowledge has shown that scientific models often function not as global representations but as selective, domain-bound, and purpose-relative structures. The contribution of the present paper accordingly does not consist in being the first to claim that scientific models are domain-bound; it consists in linking this insight with a two-stage theory of contextual and global falsification, a reconstructive utility structure, and the epistemic enabling space $E(t)$.

2.6 Approximate Truth

Niiniluoto (1987, 1998) and Oddie (1986) offer graded conceptions of truth but do not explain why some models persist despite partial falsifications while others disappear entirely. What is missing in particular is a coupling of approximate truth to contexts of use, model costs, and institutional stabilization. As a result, it remains unclear how graded degrees of truth are supposed to operate within actual model portfolios.

2.7 Research Gap

The relevant literature has already supplied many of the elements central to a theory of modern model evaluation: the underdetermination of empirical tests, the persistence of anomalies, the stability of research programs, the domain-bound character of scientific models, graded truthlikeness, model pluralism, robustness analyses, and procedures of statistical model selection. What has largely been missing, however, is an integrated framework that explicitly relates these insights to the distinction between contextual and global falsification. The research gap therefore does not lie in claiming that models are domain-specific, idealized, or pluralistically organized. It lies in the systematic linking of domain-relative model validity, relative epistemic competitiveness, model costs, available alternatives, and the epistemic enabling space $E(t)$. Such a framework must simultaneously explain:

- when falsification operates globally,
- when it remains contextual,
- how approximate truth functions within specific domains,
- how model costs and explanatory power enter into model decisions,
- how available alternatives determine the relative competitiveness of a model,
- how the epistemic enabling space $E(t)$ constrains and enables the set of real model alternatives.

The framework proposed here addresses precisely this gap. It treats falsification not as an isolated logical event but as a change in the relative validity of a model within a domain. As a result, anomalies, model persistence, approximate truth, model costs, and available alternatives are not treated separately but reconstructed as interrelated dimensions of scientific model evaluation. This is particularly relevant for data-intensive, simulation-based sciences, in which contextual falsification and model portfolios have become standard practice.

2.8 Contribution of This Paper

The present paper makes four interrelated contributions to the philosophy of science and the philosophy of models. It understands itself as a local philosophy-of-science specialization within a more general account of model management under finite conditions. Its aim is not to develop an independent metatheory alongside that account but to make the general concepts of model validity, the domain-bound character of models, costs, friction, and revision formally applicable to the special case of scientific falsification.

1. **The concept of the epistemic enabling space $E(t)$.** $E(t)$ is articulated as a dynamic structure of methodological, technical, and institutional conditions that determines which models can be formulated, evaluated, and stabilized. This brings into view a dimension that is only implicitly considered, if at all, in Popper's classical theory of falsification.
2. **The two-stage falsification structure.** The paper strictly distinguishes contextual from global falsification. Global falsification is no longer understood as a binary breach of truth but as a state in which a model, under the conditions of $E(t)$, no longer remains epistemically competitive in any relevant domain D . This structure systematically reconstructs the stability of partially falsified models.
3. **The integrated utility function $U(M, D, t)$.** With the utility function $U(M, D, t) = \alpha \cdot AT(M, D, t) + \beta \cdot EK(M, D, t) - \gamma \cdot C(M, D, t)$ the paper proposes a framework that brings approximate truth, explanatory power, and model costs together within a single evaluative structure and explicitly anchors them in $E(t)$. The function is designed to remain compatible with established measures of fit and selection criteria (fit metrics, information criteria, complexity measures) used in particular disciplines.
4. **Application to simulation-based model portfolios.** The case of climate models is used to show how ensemble structures, parameterizations, and adaptive weighting procedures can be reconstructed within the proposed framework as cases of contextual falsification and domain-specific model revision. The theory thus connects directly with a central class of contemporary, data- and simulation-intensive modeling practices, without claiming that all climate models themselves follow a uniform Bayesian updating mechanism.

3. Basic Concepts

The following definitions sharpen the concepts central to this paper. They apply general categories of model management under finite conditions to the special case of scientific falsification. Model validity, the domain-bound character of models, costs, friction, and revision are thereby not left as abstract background notions but are translated into a formally applicable structure. The quantities $U(M, D, t)$, $AT(M, D, t)$, $C(M, D, t)$, and $E(t)$ serve to render this structure explicitly reconstructible in the context of scientific model evaluation.

Notation (Overview)

M – model M^* – available alternative model $D(M)$ – domain of validity of a model $D_1, D_2, \dots, D_n$ – subdomains of the domain of validity $AT(M, D, t)$ – approximate truth of the model in domain D at time t $EK(M, D, t)$ – explanatory power of the model in domain D at time t $C(M, D, t)$ – model costs of the model in domain D at time t $U(M, D, t)$ – domain-specific utility function $\Delta U(M^*, M, D, t)$ – relative utility difference between alternative model M^* and model M $E(t)$ – epistemic enabling space at time t $R(M, D_i)$ – restricted application of a model to a subdomain D_i ϵ – context-dependent tolerance range of epistemic competitiveness $D(M)'$ – remaining domain of validity after restriction or exclusion of a problematic subdomain $Rev(M, D_i)$ – revision of a model or submodel within a problematic subdomain

Definition 1: Model (M)

A model is an idealized mathematical or conceptual structure that represents specified aspects of a target system under particular conditions (Cartwright 1999).

Definition 2: Domain of Validity $D(M)$

The domain of validity of a model designates that domain within which the model yields reliable, explanatorily powerful, or functionally viable results. A domain is not merely an external area of application but a region of stable conditions of validity: within these conditions, a model can be tested, compared, restricted, or defended against alternatives. $D(M)$ therefore designates not the global truth of a model but its domain-relative epistemic validity. This domain-relative reading connects with the theory of domains, boundaries, and transition functions, in which validity is reconstructed as dependent on the stability conditions of a particular domain (Rapp 2026e).

Definition 3: Subdomains D_1, D_2

$D(M)$ can be decomposed into a set of subdomains $\{D_1, D_2, \dots, D_n\}$. Such decompositions are often the result of scientific revision, domain refinement, or model comparison. For simple cases, one can distinguish between preserved domains and problematic domains: preserved domains are those in which M remains epistemically optimal or competitive, while problematic domains are those in which M is weakened, in need of revision, or no longer epistemically competitive against alternatives. The binary distinction between D_1 and D_2 is therefore only a simplified special case of a more general subdomain structure.

Definition 4: Epistemic Optimality

A model is epistemically optimal in a domain D if, with respect to all alternatives available in the epistemic enabling space $E(t)$, it exhibits the most favorable trade-off among approximate truth, explanatory power, and model costs. Epistemic optimality is therefore not an absolute property of a model but a relational evaluation: a model is optimal only relative to a domain, to a time t , and to the alternatives actually available within $E(t)$. Operationally, this evaluation is reconstructed by the domain-specific utility function $U(M, D, t)$.

Approximate truth designates the domain-specific fit of a model to the empirical, structural, or interventionally testable features that are relevant within that domain. Explanatory power, by contrast, designates a model's explanatory and structuring capacity: its capacity to render phenomena intelligible, to structure interrelations, to make causal or counterfactual dependencies visible, and to give theoretical orientation to a domain. Fit and explanation are thereby distinguished analytically: a model can fit certain data or structures well without yet possessing high explanatory power; conversely, explanatory power remains epistemically limited if the model's fit to the relevant features of a domain is too weak.

Model costs $C(M, D, t)$ cover three dimensions: (1) epistemic costs (e.g., susceptibility to error, sensitivity, the breadth of uncertainty in predictions, interpretability/transparency, quality of fit); (2) technical costs (computational effort, data requirements, implementation complexity); (3) institutional costs (degree of standardization, availability of infrastructure).

Some of these cost components are global (for instance, basic implementation or infrastructure costs), while others are domain-specific (for instance, data demands or numerical stability in particular domains). For the utility function, these are aggregated into $C(M, D, t)$.

Definition 5: Contextual Falsification

An observation O leads to contextual falsification if it shows that there is a subdomain D_2 in which M is no longer epistemically acceptable or competitive, while at least one subdomain D_1 remains in which M is epistemically optimal or competitive. Contextual falsification is thus a specific form of epistemic friction: it indicates the boundary of the prior model validity without eliminating the model as a whole. Its typical consequence is not global rejection but revision, restriction, or reorganization of the domain of validity. Such a reorganization is epistemically legitimate, however, only if it accomplishes more than the mere rescue of a weakened model: it must make new testable distinctions visible, separate sources of error diagnostically, refine model applications, or enable new explanatory, predictive, or interventionist achievements. In this sense, contextual falsification is closely linked to revision under finite conditions, insofar as it does not simply negate model validity but triggers reasoned adjustments, restrictions, or transformations of the domain of validity (Rapp 2026c).

Definition 6: Global Falsification

A model M is globally falsified at time t if, under the conditions of the epistemic enabling space $E(t)$, it remains neither epistemically optimal nor epistemically competitive in any relevant domain D . Epistemic optimality holds in a domain D if $U(M, D, t)$ is greater than or equal to the utility value of all alternative models available in the enabling space $E(t)$. Epistemic competitiveness holds if $U(M, D, t)$ lies below the utility value of the best alternative $U(M^*, D, t)$ by at most a context-dependent tolerance range ε .

Formally, global falsification can be expressed as follows: for every relevant domain D , there exists at least one alternative model M^* such that:

$$U(M, D, t) + \varepsilon < U(M^*, D, t).$$

The tolerance parameter ε reflects the fact that models with marginally lower utility values can still remain competitive in practice. To prevent ε from becoming a means of immunizing a weakened model, the tolerance range is not to be set arbitrarily but is to be reconstructed from the relevant disciplinary evaluation practices, error tolerances, uncertainty measures, and contexts of application. Operationally, ε tips from legitimate tolerance into immunization once it grows larger than the margins of error and uncertainty recognized in the domain in question. If the utility difference $\Delta U(M^*, M, D, t)$ systematically exceeds these disciplinary standards, it can no longer be absorbed by ε without disabling the discipline's evaluative practice itself. ε therefore designates not a protective zone for inferior models but the range within which competitiveness is still recognized as obtaining by the evaluation procedures customary in the domain. Global falsification accordingly does not arise from mere or short-term inferiority to an alternative but only from stable relative non-competitiveness across all relevant domains—that is, when the inferiority to available alternatives is not merely punctual but persists across the relevant testing and application contexts.

Supplement: Relevant Domain

A relevant domain is a scientific context of use in which a model continues to provide explanatory, predictive, interventionist, or structuring performance within current research or applied practice. Not every residual use of a model therefore grounds epistemic competitiveness. Historical, didactic, metaphorical, or purely illustrative uses can remain legitimate but do not count as relevant domains in the sense of global falsification. A model can thus be preserved historically or didactically and nonetheless be globally falsified in the epistemic sense if it is no longer competitive in any current scientific domain. The relevance of a domain is not derived solely from the utility value $U(M, D, t)$ but is reconstructed from existing scientific practice, its target quantities, testing procedures, contexts of application, and recognized evaluative orderings. This prevents domain relevance from being inferred merely from the performance evaluation of the model under examination.

Definition 7: Epistemic Enabling Space $E(t)$

The space of methodological, mathematical, institutional, and technical preconditions that determines:

- which models can be formulated,
- which data are available,
- which idealizations are admissible,
- which models can be stabilized,
- how model costs $C(M, D, t)$ can be determined at all.

$E(t)$ is dynamic: technological innovations, data availability, and institutional norms alter the space of possible models.

Definition 8: Approximate Truth

Approximate truth designates the degree of stabilized fit between model structure and the comparison features that are relevant within a domain. Such comparison features can include observational data, measurement structures, robust system features, interventionally testable results, or theoretically distinguished structures. Approximate truth is therefore neither a global property of a model nor a direct grasp of a target system as such, but a domain-specific and reconstructive evaluative quantity.

Formally, approximate truth can be expressed as a similarity-and-fit metric:

$$AT(M, D, t) = \sum_i w_i \cdot \text{sim}(M, S_i, D, t).$$

Here, S_i designates a comparison feature determined as relevant within domain D ; w_i describes its relative weighting within the relevant scientific evaluative ordering. Interventionist practice can enter as a testable fit relation, for instance when a model correctly represents stable changes under interventions. The theoretical interpretation, causal classification, or counterfactual disclosure of such relations belongs, however, to the explanatory power $EK(M, D, t)$ and is to be analytically distinguished from AT .

In practical model evaluation, approximate truth is always considered together with uncertainty estimates of the model's predictions. High predictive uncertainty reduces a model's effective epistemic value even when its mean fit is high, and can therefore affect both $AT(M, D, t)$ and the epistemic costs $C(M, D, t)$.

Definition 9: Domain Operators

The following operators describe different forms of change or refinement of model validity.

- Restriction: $R(M, D_i) = M|D_i$ designates the restricted application of a model to a particular subdomain D_i .
- Domain difference: $D(M)' = D(M) \setminus D_i$ designates the remaining domain of validity of a model after exclusion or relinquishment of a problematic subdomain D_i .
- Revision: $Rev(M, D_i)$ designates the targeted adjustment of a model or submodel within a problematic subdomain D_i .
- Decomposition: $Z(M) = \{D_1, D_2, \dots, D_n\}$ designates the partition of a model's domain of validity into several testable and evaluable subdomains.

These operators distinguish three cases that are often conflated in modeling practice: the restricted application of a model within a subdomain, the restriction of its domain of validity, and local revision within a problematic domain.

On Temporal Dynamics

Since the epistemic enabling space $E(t)$ changes in the course of scientific development, the evaluative quantities $AT(M, D, t)$, $EK(M, D, t)$, and $C(M, D, t)$ are in principle time-dependent. The time index t is used explicitly in what follows whenever the dynamics of $E(t)$, the change in available alternatives, or the historical shift of model validity are at issue. Where it would impede readability, the index t may be omitted; in substance, however, it is always to be understood as present.

4. Criteria of Modern Model Evaluation within the Proposed Framework

Modern scientific model evaluation involves a range of epistemic, technical, and institutional criteria that together determine whether a model is preserved, recalibrated, or disappears within a domain D . The framework developed here renders these criteria explicit and assigns them systematically to the components of the utility function $U(M, D, t)$ and to the epistemic enabling space $E(t)$. The following catalog does not claim to capture all discipline-specific forms of evaluation exhaustively, but rather to identify the central epistemically relevant evaluation dimensions of modern modeling practice.

(1) Approximation and structural fit

This includes domain-specific approximate truth $AT(M, D, t)$, understood as the graded fit between model structure and relevant comparison features within a domain. Components include empirical quality of fit, agreement with robust data structures, structural similarity, stability of model predictions, and, where applicable, the correct representation of interventionally testable relations. AT thus designates a model's fit dimension: it indicates how well a model fits the features that are relevant within a given domain.

(2) Explanatory power (EK)

A model's explanatory power concerns what it contributes beyond mere fit. It encompasses the capacity to render phenomena intelligible, to unify diverse findings, to make causal or interventionist dependencies visible, to enable counterfactual insights, and to give theoretical structure to a domain. $EK(M, D, t)$ is therefore not reducible to predictive accuracy, fit, or structural similarity. Two models can possess similar approximate truth and yet differ markedly in how much they explain, order, or theoretically disclose.

(3) Model uncertainty

Model uncertainty covers the variance, sensitivity, and stability of model predictions within a domain. Wide uncertainty bands diminish a model's effective epistemic value even when its mean fit is high. They affect both $AT(M, D, t)$ and the epistemic costs $C(M, D, t)$.

(4) Interpretability and transparency

Models differ in their structural accessibility, the traceability of their mechanisms, their testability, and the diagnostic possibilities they offer for tracing sources of error. Low interpretability raises epistemic costs and lowers a model's explanatory usability.

(5) Technical and computational costs

These include computational effort, energy consumption, numerical complexity, data requirements, and implementability. Such technical costs form a central component of $C(M, D, t)$ and substantially influence the rationality of model choice in data-intensive sciences.

(6) Institutional and infrastructural conditions

Models are stabilized through norms, software ecosystems, training structures, data formats, and organizational practices. These factors are located in the epistemic enabling space $E(t)$ and explain why models often persist even when they become suboptimal in particular domains.

(7) Availability of epistemic alternatives

The set of alternative models available in the enabling space $E(t)$ determines whether contextual falsification leads to model revision, domain restriction, or global elimination. Models persist as long as they remain epistemically optimal or competitive in at least one relevant domain. What matters, therefore, is not only a model's internal performance but its relative position vis-à-vis available, testable, and stabilizable alternatives.

This catalog of criteria specifies which factors carry operational significance within the utility function $U(M, D, t)$ and the enabling space $E(t)$. At the same time, it provides a systematic overview of the central evaluative dimensions that shape modern scientific modeling practice. The integrated framework is thus made theoretically applicable and empirically interpretable, without claiming to provide a complete catalog of all discipline-specific forms of evaluation.

5. Proposal: Two-Stage Falsification Structure

In the terminology proposed here, "falsification" no longer designates a binary decision about truth or falsity about M as a whole but a systematically reconstructible shift in the utility values $U(M, D, t)$ across domains, which under certain conditions leads to the complete relinquishment of M .

For every model M :

1. As long as there exists at least one relevant subdomain in which M is epistemically optimal or competitive according to $U(M, D, t)$, falsification does not lead to global elimination but to restriction or revision of the domain of validity: $D(M) \rightarrow D(M)'$, where $D(M)'$ designates the remaining relevant domains in which M is epistemically optimal or competitive.
2. M is globally falsified if and only if there is no relevant domain D in which M is epistemically optimal or competitive according to $U(M, D, t)$.

Approximate truth $AT(M, D, t)$, explanatory power $EK(M, D, t)$, and model costs $C(M, D, t)$ determine the evaluation of epistemic optimality; $E(t)$ determines the set of available and stabilizable alternatives.

Formally, the transition from contextual to global falsification can be expressed as a shift from relative competitiveness to stable non-competitiveness. Throughout, M^* designates the best alternative model available within the domain in question. Contextual falsification obtains as long as there exists at least one relevant domain D in which M remains competitive within the tolerance range ε :

$$\exists D: U(M, D, t) + \varepsilon \geq U(M^*, D, t).$$

Global falsification, by contrast, obtains when M lies stably below this tolerance range across all relevant domains:

$$\forall D: U(M, D, t) + \varepsilon < U(M^*, D, t).$$

The transition is therefore not a mere switch from truth to falsity but a change in the relative validity of a model with respect to available alternatives.

This two-stage structure makes formally tractable a distinction that, while discussed in the literature, has rarely been modeled explicitly. It precisely reconstructs how models can be preserved in particular subdomains even though they fail in others, and thereby supplements the existing accounts of Popper and Lakatos with a clearly structured representation of domain-specific model stability. The two-stage falsification structure thus describes explicitly the situation, frequently observed in scientific practice, in which models are recalibrated in subdomains while their overarching epistemic role is preserved.

5.1 Limiting Routes of Immunization

The theoretical robustness of the two-stage falsification structure depends on the exclusion of three systematically connected routes of immunization. First, the tolerance range ϵ must not become an open-ended protective clause for weakened models. It designates only the zone within which competitiveness remains plausible according to the standards of error, uncertainty, and evaluation recognized in a domain. Second, the relevance of a domain must not be derived from the defensive interest in the model under examination. A relevant domain must be stabilized independently through existing scientific practice, target quantities, testing procedures, contexts of application, or recognized evaluative orderings. Third, domain decomposition must not function as an ad hoc rescue. It is epistemically legitimate only if it makes new testable distinctions visible, separates sources of error diagnostically, refines model applications, or enables new explanatory, predictive, or interventionist achievements.

These three constraints work together. A theory of contextual falsification that draws only one of these limits remains open to immunization through the others: through an overly generous ϵ , through artificially generated residual domains, or through merely defensive subdomain construction. Contextual falsification is therefore rationally reconstructible only when tolerance ranges, domain relevance, and domain decomposition are each constrained by external disciplinary standards, existing forms of practice, and new epistemic gains. Thus, for instance, a model whose tolerance range ϵ is well constrained at the disciplinary level could still be preserved by a merely defensive subdomain construction unless that subdomain construction is supported by independent practice standards or by new epistemic gains.

6. When Models Disappear

A model does not disappear merely because it shows weaknesses in particular domains or performs marginally worse than an alternative. Global model replacement requires that, under the conditions of $E(t)$, M remains neither epistemically optimal nor competitive in any relevant domain. What is decisive, therefore, is stable relative non-competitiveness against available alternatives. This diagnosis presupposes that ε is not artificially expanded, that domain relevance is not derived merely from the defensive interest in the model, and that domain decomposition is not used as a purely defensive rescue strategy.

A model typically disappears when:

1. across all relevant domains, at least one alternative model M^* durably enjoys a utility advantage beyond the tolerance range ε : $U(M, D, t) + \varepsilon < U(M^*, D, t)$,
2. the model costs of M relative to its alternatives are no longer compensated for by approximate truth or explanatory power,
3. under the conditions of $E(t)$, there are no current scientific contexts of use or relevant domains in which M remains epistemically competitive.

Examples:

- ether → Maxwell (1865) and Einstein (1905),
- phlogiston → modern chemistry,
- epicycles → Kepler (1609, 1619) and Newton,
- the four humors theory → modern medicine.

These historical model changes are often interpreted in the philosophy of science as paradigm shifts or transformations of research programs (Kuhn 1962; Lakatos 1970), but they receive a more precise formal reconstruction within the utility structure proposed here.

Technological, mathematical, and institutional innovations alter $E(t)$ and accelerate global falsification. $E(t)$ thereby determines not only which models disappear but also which can even count as realistic alternatives.

7. Why Models Persist

7.1 Robustness through Model Families

Many models exist not as individual structures but as entire model families. Falsification therefore typically targets variants or particular parameterizations rather than the entire model class. Model families possess structural redundancies that allow errors in some submodels to be compensated for without abandoning the overall approach. This explains why falsification often leads only to local revision, internal reweighting, or restriction of the domain of validity, rather than to elimination of the entire class. This observation is consistent with the model-theoretic literature, in which model families are described as structured spaces of possible variants (Weisberg 2013). The domain structure introduced here refines this approach by showing how stability and variation within a model class are formally interdependent.

7.2 Contexts of Use

Models persist for different reasons. Some of these reasons ground epistemic competitiveness within relevant domains; others merely explain historical, didactic, or institutional persistence.

Models can remain scientifically relevant because they:

- are explanatorily powerful or predictively useful in particular domains,
- remain computationally efficient,
- can be implemented reliably in technical terms,
- function as stable building blocks within larger model architectures.

In addition, models can persist didactically, historically, or illustratively. Such residual uses explain their cultural or institutional persistence but do not, by themselves, ground a relevant domain in the sense of global epistemic competitiveness. Technical standardization can reinforce this persistence further, but does not on its own ground epistemic competitiveness.

Contextual falsification corrects domains but does not eliminate the model. The persistence of models in D_1 despite falsifications in D_2 thus follows from the utility structure $U(M, D_1, t)$ and from the stabilization of institutional and technical contexts within the epistemic enabling space $E(t)$. The existence of stable contexts of use is a function of the epistemic enabling space $E(t)$: institutions, data formats, software libraries, and training structures secure models independently of their global approximate truth. This explains why models whose utility structure is weakened in a subdomain D_2 can still be used rationally in another domain D_1 . This corresponds to a practice observable in many sciences, in which models function as tool-like building blocks whose validity is context-dependent and not assessed globally.

8. Case Study: Climate Models

8.1 Climate Models as Ensemble Structures

Climate models are not individual models but complex ensembles of model variants and scenarios (Oreskes et al. 1994):

$$\text{Ensemble} = \{M_1, M_2, \dots, M_n\}$$

This ensemble structure allows the combination of different physical assumptions, parameterizations, and initial conditions.

Falsification therefore rarely targets the entire ensemble but specific components:

- ocean models,
- atmospheric parameterization,
- cloud processes,
- biogeochemical modules.

The ensemble as a whole remains epistemically stable even when individual components are modified or replaced. This corresponds to established modeling practice in climate science, in which ensembles function explicitly as mechanisms of epistemic stabilization (cf. Oreskes et al. 1994; IPCC 2021). The domain structure of the ensemble is therefore more stable than that of the individual models.

8.2 Parameterization as an Epistemic Operation

Many climate-relevant processes cannot be computed at full resolution from fundamental physical equations. Parameterizations serve as epistemic bridges between physical theory and numerical feasibility.

Falsification typically marks the limits of such parameterizations. This leads to local domain adjustments in which the overall model is not necessarily rejected; instead, the utility structure of a submodel or parameterization decreases within a particular subdomain:

$$U(M, D_2, t) \downarrow \rightarrow \text{Rev}(M, D_2) \text{ or } D(M)' = D(M) \setminus D_2.$$

Parameterizations are therefore not merely techniques of approximation but operational definitions of domains, the adjustment of which is a central mechanism of contextual falsification. The overall approach is preserved as long as the model or model family remains epistemically competitive in other relevant domains.

8.3 Bayesian Updating and Adaptive Modeling

Bayesian procedures can be employed in evaluating, weighting, and combining climate models—for example, in the context of model weighting, ensemble evaluation, or probabilistic projection:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

This formula does not assert that climate models in general follow a uniform Bayesian updating mechanism. It rather stands ideal-typically for a class of redistributive evaluation procedures in which empirical deviations do not necessarily lead to elimination of an overall model but can change the relative weighting of submodels, parameterizations, or model variants.

A poorly predicted variable can reduce the likelihood of a submodel or parameterization and thereby weaken its domain-specific approximate truth $AT(M, D_2, t)$. This does not necessarily lead to elimination of the entire ensemble. Rather, alternative parameterizations, model variants, or submodels within the model portfolio can be weighted more heavily if they achieve higher epistemic competitiveness in the relevant domain.

At the same time, caution is warranted in actual climate modeling. The philosophy of climate modeling has repeatedly shown that ensemble agreement cannot be translated immediately into epistemic certainty. Model weightings are contested because climate models are often not statistically independent, share common model genealogies, similar parameterizations, or correlated error structures. Robustness across multiple models is therefore strongly load-bearing only when it does not merely reproduce the same background assumptions, data structures, or sources of error but is supported by sufficiently independent lines of evidence, model comparisons, or process understanding. Precisely for this reason, climate modeling is a particularly suitable case for the theory developed here: it shows that model validity cannot be determined by fit or ensemble agreement alone, but only through the interplay of approximate truth, explanatory power, uncertainty, model costs, available alternatives, and $E(t)$. Bayesian updating is therefore introduced here not as a universal procedure for climate model evaluation but as a formal illustration of how falsification can operate redistributively within complex model portfolios. What matters for the theory developed here is that such shifts can be reconstructed as a domain-specific change in relative competitiveness within a model portfolio, regardless of whether the discipline in question employs Bayesian, frequentist, heuristic, or ensemble-based evaluation procedures.

8.4 Domain Structure

The functional areas of climate models can be ideal-typically organized as follows:

- D_1 : global temperature trends,
- D_2 : regional precipitation patterns,
- D_3 : extreme events,
- D_4 : short-term climate variability, such as ENSO or AMO.

These domains differ in data availability, spatial and temporal resolution, breadth of uncertainty, and predictive stability. In many evaluative contexts, global temperature trends are taken to be more robustly reconstructible than regional precipitation patterns, extreme events, or short-term variability. From this, however, no simple hierarchy of absolute model validity follows. Rather, the epistemic viability of climate models must be evaluated in domain-, scale-, and scenario-dependent terms. This is precisely why climate modeling is well suited as an example of contextual falsification: deviations in individual subdomains do not necessarily weaken the entire model family but often lead to local revision, reweighting, or domain refinement.

8.5 Miniaturized Formal Illustration

A simplified example may help clarify the structure. For ease of presentation, the time index t is omitted in this miniature.

Suppose two models M and M^* are available and are compared in two domains D_1 and D_2 .

$AT(M, D_1) = 0.9$, $EK(M, D_1) = 0.8$ $AT(M^*, D_1) = 0.7$, $EK(M^*, D_1) = 0.8$ $AT(M, D_2) = 0.4$, $EK(M, D_2) = 0.5$ $AT(M^*, D_2) = 0.8$, $EK(M^*, D_2) = 0.9$

For moderate costs $C(M) \approx C(M^*)$, the result is:

- In D_1 , M remains epistemically optimal.
- In D_2 , M^* becomes optimal.
- Globally, no elimination of M follows so long as M remains epistemically optimal or competitive in D_1 .

This illustrates the principle:

Contextual falsification weakens the utility structure of a model in a particular domain D_2 without necessarily destroying its validity in D_1 or its role within a model portfolio. This structure is characteristic of many simulation-based sciences, not only climate science.

8.6 Consequences for Epistemic Status

Climate models are often assessed according to the Popperian schema:

"An incorrect prediction shows that the model is wrong."

From a philosophy-of-science standpoint, this is incorrect (Oreskes et al. 1994). The correct view is that falsification operates in a domain-specific manner and indicates which subcomponents must be developed further. This makes clear that the Popperian elimination logic is inadequate for simulation-based model architectures and must be replaced by a domain-specific utility structure.

The case study of climate models serves as an illustrative application of a modern, data- and simulation-intensive model architecture. It shows by example how ensemble structures, parameterizations, and adaptive weightings can be reconstructed as cases of contextual falsification and domain-specific model revision. The structure of the analysis is in principle transferable to other model architectures, including classical physical models, economic models, epidemiological models, and AI model architectures. This transferability is to be understood, however, as theoretical applicability, not as completed empirical examination across all those areas.

9. Approximate Truth

Approximate truth explains why models can generate epistemic progress despite errors. The concept was developed especially by Niiniluoto (1987, 1998) and Oddie (1986), but it is used here in an explicitly domain-specific and reconstructive manner. Approximate truth in this paper does not designate the immediate access of a model to a target system as such, nor the entire epistemic performance of a model. It designates a particular evaluative dimension within $U(M, D, t)$: the domain-specific fit between model structure and the relevant empirical, structural, or interventionally testable comparison features.

Truthlikeness is thus understood not as a global property of a model but as a functionally reconstructible relation within particular conditions of validity. Approximate truth is therefore narrower than overall epistemic optimality. A model can have a strong fit to certain data or structures and yet possess limited explanatory power; conversely, a theoretically powerful model can be epistemically weakened if its fit to the relevant features of a domain is inadequate.

Formally, approximate truth can be expressed as a domain-specific similarity-and-fit metric:

$$AT(M, D, t) = \sum_i w_i \cdot \text{sim}(M, S_i, D, t).$$

Here, S_i designates a comparison feature determined as relevant within domain D , such as observed system states, empirical data structures, theoretically distinguished properties, or interventionally testable target quantities. The function $\text{sim}(M, S_i, D, t)$ describes the similarity or fit between model M and the corresponding comparison feature S_i under the methodological and empirical conditions available at time t . The weights w_i represent the relative relevance of these features within the relevant scientific evaluative ordering. Within this formal framework, approximate truth is neither an absolute criterion of truth nor a mere measure of fit; it is a domain-specific measure of structural and empirical fit that is central to model choice.

The important points are:

1. Approximate truth is always domain-specific.
2. Approximate truth designates the fit dimension of a model, not its overall epistemic performance.
3. Explanatory power remains analytically distinct: it concerns the explanatory, unifying, causal, counterfactual, or structuring performance of a model.
4. Model costs likewise remain distinct: they concern the epistemic, technical, and institutional efforts, burdens, or dependencies of a model.
5. Falsification can lower $AT(M, D_2, t)$, weaken $EK(M, D_2, t)$, or raise $C(M, D_2, t)$; but it need not automatically destroy the validity of M in other domains D_1 .
6. Global falsification cannot be defined through approximate truth alone. What matters is the entire utility structure and the relative competitiveness against available alternatives: $U(M, D, t) + \varepsilon < U(M^*, D, t)$ for all relevant domains D .

The coupling of approximate truth to the utility function $U(M, D, t)$ shows that graded truthlikeness does not operate in isolation but, in interplay with explanatory power and model costs, determines epistemic optimality. Approximate truth operates within domains and supplements the theory with a quantitative dimension of epistemic improvement.

For practical application, it is important that both the similarity function $\text{sim}(M, S_i, D, t)$ and the weights w_i are determined in a context-dependent manner and can be operationalized differently across disciplines. In many scientific areas, $\text{sim}(M, S_i, D, t)$ can be approximated through established fit-and-error metrics, including likelihood functions, variance measures, prediction errors, residual analyses, or measures of similarity between model trajectories and observed system states. In other contexts, especially in theoretically or structurally oriented sciences, similarity metrics additionally capture qualitative or topological features such as the preservation of symmetries, invariance structures, or the reproduction of causal dependence patterns.

The weights w_i represent the relative relevance of the various comparison features selected within a domain. They can arise from empirical evaluation practices, disciplinary standards, or model-theoretic considerations. Approximate truth is therefore not tied to a single metric but constitutes an adaptable structure that takes account of both the numerical accuracy and the structural fit of a model, without presupposing direct access to a target system as such.

10. Why an Integrated Theory Has Been Lacking

10.1 Historical Reasons

Before the digital era, many scientific models were formulated analytically, so domain structures were less visible. Only numerical simulations and large-scale data models made contextual falsification a central component of scientific practice.

10.2 Logical Reasons

Popper's approach was formulated primarily as a methodological principle of demarcation and testing. For analyzing modern modeling practice, however, the binary elimination structure often derived from it appears too coarse. Modern models are frequently approximative, domain-specific, and dynamically recalibrable. A purely binary understanding of falsification is unsuitable here, because approximate truth can vary across domains and empirical deviations need not concern the model as a whole. This shows that classical logical structures of falsification capture the operational complexity of modern modeling practices only to a limited extent so long as they do not explicitly represent graded approximation, the domain-bound character of models, and weightings of relevance.

10.3 Institutional Reasons

Scientific practice is embedded in:

- computational capacities,
- data norms,
- peer-review structures,
- research funding,
- technological infrastructures.

These factors shape the epistemic enabling space $E(t)$ and thereby determine which models can be formulated and stabilized. Popper's approach does not explicitly model this dimension. By contrast, in the theory introduced here $E(t)$ is explicitly modeled as the structuring force of institutional and technical conditions and thus constitutes a central epistemic variable.

11. The Epistemic Enabling Space

11.1 Dynamics of the Epistemic Enabling Space

The epistemic enabling space $E(t)$ changes with technical, methodological, and institutional development. It is not only an enabling space but at the same time a constraining space: it determines which models can be formulated, tested, compared, and stabilized, but it likewise constrains which alternatives become visible, applicable, or institutionally acceptable at all. A model can be epistemically optimal at one time and no longer so at a later time without any change to its internal structure. What is decisive is that $E(t)$ itself changes: new data, methods, computational capacities, or standards shift the set of available alternatives and thus the relative validity of models. At the same time, successful models feed back into $E(t)$ by establishing data formats, computational infrastructures, institutional standards, and research practices. A reciprocal dynamic thus emerges: $E(t)$ enables and constrains model choice, while stabilized models themselves help shape the future enabling space.

The principal influencing factors are:

1. data growth and improved measurement systems,
2. new mathematical methods,
3. increasing computational power,
4. processes of institutional standardization,
5. the development of scientific discourses.

These factors can be construed as operators of change $\Delta E(t)$ that systematically expand or restrict the epistemic enabling space.

$E(t)$ structures:

- the set of actually available models,
- the structure of their domains,
- their costs $C(M, D, t)$,
- the available alternatives M^* .

The epistemic enabling space is therefore the central frame parameter of model choice. From the perspective of the history of science, the concept $E(t)$ explains why model changes are often triggered by technological and institutional innovations that first make new model alternatives formalizable.

11.2 Structure of the Epistemic Enabling Space $E(t)$

For more precise analysis, the epistemic enabling space $E(t)$ can be decomposed into three functional subcomponents that jointly determine which models can be formulated and stabilized.

1. **$E_m(t)$: methodological-mathematical conditions.** These include available mathematical procedures, statistical methods, modeling approaches, and algorithmic techniques. They determine which types of models can be formulated at all and which approximations are admissible.
2. **$E_t(t)$: technical and data-related conditions.** These include computational power, data quality, software infrastructures, numerical tools, and simulation technologies. These factors determine which models are practically implementable and at what resolution, stability, or complexity modeling can be carried out.
3. **$E_i(t)$: institutional and organizational conditions.** This dimension covers scientific norms, peer-review procedures, funding logics, training structures, and established research practices. It shapes which models are stabilized in the long term, which standards become entrenched, and which alternatives count as scientifically acceptable.

$E(t)$ arises from the interplay of these three dimensions. The availability and stabilization of a model thus result not only from its approximate truth, explanatory power, or costs but also from the methodological, technical, and institutional conditions that enable or constrain its formulation, implementation, and further development. At the same time, $E(t)$ determines which alternative models can even appear as relevant comparison models. $E(t)$ thereby influences not only the absolute evaluation of a model but also its relative epistemic competitiveness against available alternatives.

12. The Decision Logic of Model Choice

12.1 Structure of the Utility Function

The choice between two models M and M^* can be reconstructed through a time- and domain-specific utility function:

$$U(M, D, t) = \alpha \cdot AT(M, D, t) + \beta \cdot EK(M, D, t) - \gamma \cdot C(M, D, t).$$

This function describes not the absolute value of a model but reconstructs its domain-specific epistemic validity under finite conditions. It is neither a first-order measurement formula nor an algorithmic decision procedure. Its task is to render existing disciplinary evaluation practices explicit within a common comparative structure. In this sense, it functions as a reconstructive evaluative grammar of the second order: it organizes the criteria by which models are in fact assessed within a domain, but it does not replace those criteria with a universal computational formula.

What is decisive, therefore, is not the isolated utility value of a model but its difference relative to alternatives available within the same relevant domain. This difference can be formulated as a comparative structure:

$$\Delta U(M^*, M, D, t) = U(M^*, D, t) - U(M, D, t).$$

ΔU is not an additional independent theory but a reconstructive comparative structure. It shows whether an alternative becomes only marginally better, distinctly superior, or epistemically dominant within a domain. ΔU thereby connects the utility function with the question of when contextual falsification leads to revision, domain restriction, or global model replacement.

A global evaluation can be obtained through an aggregate function $U(M)$ that weights the domain-specific values $U(M, D, t)$ across the relevant domains. Such aggregation, however, remains secondary to the domain-specific evaluation, since model validity is determined within this framework primarily within particular domains.

The components are:

- $AT(M, D, t)$: domain-specific approximate truth, that is, the degree of stabilized fit between model structure and relevant empirical, structural, or interventionally testable comparison features within a domain.
- $EK(M, D, t)$: explanatory power of the model in domain D , that is, its capacity to explain phenomena, make relations visible, enable causal or counterfactual insights, and structure a domain theoretically.
- $C(M, D, t)$: model costs, including data demands, computational effort, interpretability, institutional infrastructure, implementation effort, uncertainty, susceptibility to error, and dependence on stabilizing background conditions.
- α, β, γ : weighting factors that are not to be understood as arbitrary stipulations by individual actors but as reconstructible weightings of disciplinary evaluative orderings within the epistemic enabling space, in particular its methodological and institutional dimensions $E_m(t)$ and $E_i(t)$.

Different sciences weight approximate truth, explanatory power, and model costs differently, depending, for instance, on whether predictive accuracy, mechanistic explanation, computational efficiency, institutional standardization, or practical applicability is paramount. The parameters α , β , and γ render these evaluative orderings explicit without inventing them.

A simple example brings this reconstructive logic into view. In a prediction-oriented climate domain, α may carry especially high weight because empirical fit, predictive stability, and uncertainty control are central evaluative quantities. In a theoretically and structurally oriented domain, β may matter more if the chief task is to unify phenomena, make causal dependences visible, or theoretically open up a previously unstructured area of phenomena. In technical real-time or applied contexts, γ may be especially relevant because computational costs, data demands, robustness, implementability, and institutional applicability significantly constrain actual model choice. α , β , and γ are therefore not freely invented parameters but condense the target quantities, testing procedures, and evaluation practices already operative within a domain into an explicit comparative structure.

The utility structure is compatible with existing procedures of model selection without being reducible to them. Information criteria such as AIC and BIC operationalize, in a limited way, trade-offs between empirical fit and model complexity; Bayes factors and Bayesian model averaging provide probabilistic variants of model comparison and weighting. Severe testing and robustness analyses likewise perform specific forms of scientific model evaluation by examining the severity of tests or the stability of results across model variants. The added value of the utility structure proposed here does not consist in replacing such procedures. $U(M, D, t)$ rather reconstructs the broader decision context within which these procedures take effect: Which domain is being evaluated? Which alternatives are actually available within the epistemic enabling space $E(t)$? Which model costs are relevant? Which explanatory performance counts within the domain in question? And when does local inferiority lead only to contextual falsification, and when to global model replacement? The utility function is therefore not a competing procedure to model selection but a framework for reconstructing the epistemic embedding of such procedures.

Many terms of the utility function can be approximated in actual sciences by empirical proxy metrics or established quantifying procedures, such as fit metrics, information criteria, complexity measures, likelihood structures, prediction errors, or sensitivity analyses. These procedures each operationalize particular partial aspects of $AT(M, D, t)$, $EK(M, D, t)$, or $C(M, D, t)$ within a given discipline. The utility function is therefore to be understood as an ideal-typical reconstructive framework: it shows how various evaluative dimensions interact, without claiming that epistemic validity can be calculated completely or context-free.

The linear form of the function serves as a heuristic minimal form. It is deliberately kept simple in order to bring the basic structure of trade-off into view. In concrete applications, however, nonlinear, threshold-based, or product-form evaluation structures may be more appropriate. This is the case, for example, when explanatory power becomes epistemically relevant only above a minimum level of approximate truth, when model costs dominate above certain thresholds, or when high uncertainty constrains a model's utility not merely gradually but structurally. The linear representation is therefore not to be understood as the claim that scientific rationality is in fact additively computable, but as an applicable starting form of a reconstructive model evaluation.

Contextual falsification can lower $AT(M, D_2, t)$, weaken $EK(M, D_2, t)$, or raise $C(M, D_2, t)$, thereby decreasing $U(M, D_2, t)$ without necessarily impairing $U(M, D_1, t)$. From a procedural perspective, this change appears as a condensation of epistemic friction: the prior model stabilization produces increasing strain, declining explanatory power, rising costs, or increasing dependence on auxiliary assumptions, without thereby automatically issuing in global elimination. This friction becomes decision-relevant when, against available alternatives, it manifests as stable relative inferiority in ΔU . Falsification thus does not operate globally automatically but first changes the domain-specific utility structure of a model.

This structure also explains why model families can remain robust. If $U(M_1, D_2, t)$ decreases, other models M_2 or M_3 within the same family may continue to achieve higher values in that domain. Falsification thereby becomes a mechanism of internal reallocation of epistemic weights, not a simple elimination logic in the Popperian sense.

12.2 Scaling and Empirical Anchoring of the Utility Function

So that the proposed utility function $U(M, D, t)$ does not remain merely theoretical, it can be approximated in many scientific contexts by established evaluation procedures. Calibration is relative: within a domain D , models are not measured against an absolute standard but compared in relation to available alternatives and to the quality criteria recognized within that domain.

Such quality criteria, depending on the discipline, may include fit metrics, error rates, likelihood structures, information criteria, predictive quality, sensitivity analyses, complexity measures, or interpretability requirements. They each operationalize, however, only certain partial aspects of $AT(M, D, t)$, $EK(M, D, t)$, or $C(M, D, t)$. The utility function therefore does not provide a context-free measurement of epistemic validity but a structured reconstruction of disciplinary evaluation practices.

Empirical anchoring thus consists not in prescribing a uniform scale for all sciences but in systematically relating existing evaluation procedures within a domain to one another. In this way, it becomes visible whether a model is stabilized, revised, restricted to particular domains, or replaced by alternatives. The theory thereby has a descriptive and weakly normative dimension: it describes actual evaluation practices and at the same time renders explicit the conditions under which model decisions are rationally intelligible.

12.3 Diagnostic Protocol of Contextual and Global Falsification

The proposed theory yields a diagnostic protocol for scientific model evaluation. It serves not as a mechanical decision procedure but as a structured application schema by which the implicit evaluative dimensions of scientific model choice can be rendered explicit. The protocol is intended to clarify whether a model remains stable, is locally revised, restricted to particular domains, reweighted within a model family, or globally replaced.

1. **Identify the model M .** It must first be clarified what is being evaluated: an individual model, a model family, a parameterization, a submodel, or a particular application of a model. This distinction is decisive because falsification often does not affect the entire model but only a variant or a specific area of use.
2. **Identify the relevant domain D .** It must then be determined within which domain the model is being tested. A domain is not merely a topic or area of application but a region of stable conditions of validity. This includes target quantities, data availability, admissible idealizations, expected accuracy, testing procedures, and practical requirements. Without a domain specification, it remains unclear whether a deviation indicates a local problem or a global failure.
3. **Reconstruct the epistemic enabling space $E(t)$.** The next step is to examine which methodological, technical, and institutional conditions obtain at time t . This includes available data, measurement procedures, computational capacities, mathematical methods, software infrastructures, disciplinary standards, and accepted evaluation practices. $E(t)$ determines not only which models are possible but also which alternatives become visible, comparable, and stabilizable at all.
4. **Capture AT , EK , and C separately.** The evaluation should first reconstruct approximate truth, explanatory power, and model costs separately. AT concerns the domain-specific fit between model structure and relevant empirical, structural, or interventionally testable comparison features. EK , by contrast, concerns the explanatory performance of the model—its ability to render relations visible, to structure relevant phenomena, and to enable, where appropriate, unifying, causal, counterfactual, or interventionist insights. C covers the epistemic, technical, and institutional costs of a model, including uncertainty, complexity, data demands, interpretability, computational effort, or infrastructural dependence.
5. *Identify alternatives M .** Falsification becomes decision-relevant only when available alternatives exist. It is therefore necessary to clarify which alternative models within the same domain are actually available, testable, implementable, and institutionally stabilizable. Determining such alternatives connects with the question of efficient search under finite conditions, since not all logically conceivable alternatives but only available, testable, and stabilizable alternatives within $E(t)$ are decision-relevant (Rapp 2026d). A model can remain stable despite weaknesses if no alternative within $E(t)$ achieves higher epistemic competitiveness.

6. **Reconstruct the relative difference ΔU .** The next question is not only whether M has weaknesses but how large the difference to M* is. ΔU shows whether an alternative is only marginally better, distinctly superior, or stably epistemically dominant across the relevant testing and application contexts. Small differences can lie within the tolerance range ϵ ; pronounced and stable differences, by contrast, can ground revision, domain restriction, or model replacement.
7. **Make the diagnosis.** The diagnosis follows from the previous steps. If M remains competitive within D, stabilization obtains. If M loses performance only in a subdomain D_2 , contextual falsification with a need for revision obtains. If M remains viable in D_1 but not in D_2 , domain restriction follows. If a submodel within a model family is weakened and another is strengthened, internal reweighting obtains. If M stably loses competitiveness across all relevant domains vis-à-vis available alternatives, global model replacement obtains. Before any of these diagnoses is finally rendered, it must be checked that ϵ remains within disciplinary tolerance standards, that the underlying domains are stabilized independently of the defense of the model, and that any domain decompositions yield progressive epistemic gains. Only when these three constraints are observed is the diagnosis rationally reconstructible.

This diagnostic protocol translates the formal structure of the paper into an applied practice. It treats falsification not in isolation as an observational contradiction but as a diagnostic process in which model validity, domain, available alternatives, costs, and enabling conditions are jointly reconstructed. The theory thereby does not become an algorithmic decision procedure, but it becomes an operative framework that renders actual model decisions more transparent and more comparable.

13. Relation to Scientific Realism

The integrated theory stands between structural realism and instrumentalist model pluralism. This intermediate position is especially relevant for modern modeling practices, since many sciences operate neither in fully realist nor in fully instrumentalist fashion. Within this framework, models possess truthlikeness insofar as they successfully capture and reconstruct stable relational, empirical, or interventionally relevant structures within particular domains. This approach is compatible with forms of structural realism, since it locates truthlikeness not in the model as a whole but in those stable relational structures that are preserved within specific domains. Their continuance, however, also depends on practical factors: costs, institutional stability, available alternative models, and shifts in the enabling space $E(t)$.

This approach supplements structural realism by rendering explicit that approximate truth is not the sole determinant of scientific rationality. At the same time, it extends instrumentalist model pluralism by formalizing when models remain epistemically optimal despite partial falsifications.

The approach developed here can be located within the realism debate as a mediating position. It is close to structural realism, in the tradition of Worrall (1989) and Psillos (1999), insofar as approximate truth is interpreted as an approximation to stable relational structures. At the same time, the approach pursues a broader perspective than van Fraassen's constructive empiricism (1980), since it shows that the operative rationality of scientific modeling practice depends not solely on empirical adequacy but also on institutional, technical, and explanatory conditions that are systematically taken into account in the epistemic enabling space $E(t)$.

In contrast to da Costa and French's conception of partial truths (2003), the present framework integrates approximate truth not only as a logical structure of graded truth but embeds it in an explicit utility function $U(M, D, t)$ that includes explanatory power and model costs. Scientific rationality thereby becomes reconstructible as a decision problem within a dynamic enabling space. The approach therefore positions itself between structural realism and instrumentalist model pluralism, but combines their insights into an integrative framework that connects directly with the empirical modeling practice of contemporary sciences.

Especially in data-based and simulation-intensive sciences—climate research, economics, epidemiology, or AI modeling—this framework provides a precise way to analyze model stability and model change systematically.

14. Falsification as Domain-Relative Revision of Validity

Falsification is not a binary mechanism of elimination but a domain-specific tool for the refinement of scientific models. The theory developed here gathers central insights from Popper, Kuhn, Lakatos, da Costa/French, and contemporary model theory within an explicitly formalized, unified framework. The distinction between global and contextual falsification shows that empirical deviations predominantly lead to a restriction of the domain of validity, not to a complete relinquishment of the model.

Integrating approximate truth, explanatory power, and model costs into a domain-specific utility function shows that model stability is the result of relative epistemic optimality and does not follow from truthlikeness alone. Models disappear only when, under the conditions of $E(t)$, they no longer remain epistemically competitive in any relevant domain or current scientific practice of use. Global model replacement is therefore not the immediate consequence of an individual empirical contradiction but the result of stable relative non-competitiveness against available alternatives. At the same time, the theory of contextual falsification is robust only if ϵ , domain relevance, and domain decomposition are not used as instruments of immunization. Falsification is thereby not weakened but refined: it constrains, revises, reweights, or eliminates model validity wherever scientific practice provides sound reasons for doing so.

With the concept of the epistemic enabling space, it also becomes possible to explain why technological, institutional, and methodological changes profoundly influence scientific model landscapes. Model choice thereby becomes reconstructible as a dynamic process in which scientific rationality consists not in a singular logical operation but in a structured interplay of context, approximation, and institutional stabilization.

The framework developed here is itself to be understood as a model. Its primary domain is the analysis of scientific model validity under finite conditions, particularly in idealized, simulation-based, and pluralistic sciences. Its own validity is therefore likewise domain-specific: it can have high reconstructive power in certain scientific contexts without being a universal theory of scientific rationality. Its costs lie in abstraction, formalization effort, and the danger of possible spurious precision. The present theory thus is itself subject to the conditions it describes: it can be contextually restricted, further refined, or supplemented by more suitable frameworks.

Falsification is thereby not abolished but refined: it remains a central instrument of rationality in scientific practice, but it operates not as a simple binary mechanism of exclusion but as a domain-relative indicator of bounded model validity. Its epistemic function is to constrain, revise, reweight, or replace model validity wherever available alternatives durably achieve higher epistemic competitiveness.

Local Philosophy-of-Science Glossary for This Paper

The following concepts stand within a modular relationship inside the Epistemics program. Friction (Rapp 2026b), revision (Rapp 2026c), search (Rapp 2026d), and domain structure (Rapp 2026e) are developed in dedicated works; the present paper transfers this architecture to the special case of scientific falsification and supplements it with the two-stage falsification structure and the epistemic enabling space. The following ordering of concepts serves to stabilize central meanings within this text. It does not claim completeness and does not establish an independent general canon of concepts. Concepts not specifically listed here are used in the sense of the Epistemics base canon or are not central to the functional core of this paper. Changes, refinements, or extensions are possible but must be explicitly indicated, locally bounded, and justified. Implicit shifts of meaning, silent extensions, or retroactive reinterpretations are excluded.

Adoption of the Epistemics Base Canon

This paper adopts the conceptual canon defined in the Epistemics base paper as an unchanged reference base (Rapp 2026a). The concepts introduced there are used without reinterpretation and without implicit shifts of their functional meaning. This paper introduces no divergent definitions of the adopted canonical concepts.

Local Philosophy-of-Science Refinements

This paper introduces no independent general extension of the canon. It refines several adopted concepts locally for the special case of scientific falsification and modern model evaluation. These refinements do not alter the Epistemics base canon but translate parts of its vocabulary into a philosophy-of-science language for describing model validity, domain restriction, model revision, and global model replacement.

Model Validity

- Short definition: domain-relative epistemic viability of a model under particular methodological, technical, and institutional conditions.
- Function: designates here the question in what area a model remains scientifically usable, testable, and competitive against alternatives.
- Demarcation: not the global truth of a model; not unrestricted validity; not mere factual or didactic use.

Contextual Falsification

- Short definition: domain-specific weakening or restriction of model validity, without elimination of the model as a whole.
- Function: describes cases in which a model is revised or restricted in a subdomain while remaining epistemically viable in other relevant domains.
- Demarcation: not a complete refutation of the model; not a mere anomaly without consequence; not an automatically global model replacement.

Global Falsification

- Short definition: loss of epistemic competitiveness across all relevant domains under the conditions of the epistemic enabling space $E(t)$.
- Function: marks the case in which a model no longer possesses any relevant scientific viability against available alternatives.
- Demarcation: not a mere partial falsification; not didactic or historical non-use; not a simple consequence of an individual observational contradiction.

Epistemic Enabling Space $E(t)$

- Short definition: the structure of methodological, technical, and institutional conditions within which models can be formulated, tested, compared, stabilized, and within which alternatives become available.
- Function: explains why model validity is historically variable and why alternatives become visible, available, and stabilizable only under particular conditions.
- Demarcation: not a neutral background space; not a mere collection of external factors; not a mechanical determination of scientific model choice.

Approximate Truth $AT(M, D, t)$

- Short definition: domain-specific fit between model structure, observational data, interventionist practice, and relevant comparison features at time t .
- Function: serves to reconstruct graded truthlikeness within particular conditions of validity.
- Demarcation: not a global truth property of a model; not direct access to a target system as such; not a mere fit metric.

Explanatory Power $EK(M, D, t)$

- Short definition: a model's capacity to render phenomena intelligible, structure them, and disclose them counterfactually within a domain.
- Function: supplements approximate truth with the explanatory and structuring performance of a model.
- Demarcation: not mere predictive accuracy; not rhetorical plausibility; not domain-independent explanatory force.

Model Costs $C(M, D, t)$

- Short definition: epistemic, technical, and institutional efforts or burdens associated with the formulation, application, stabilization, or revision of a model.
- Function: shows that model choice depends not solely on truthlikeness or explanatory power but also on uncertainty, complexity, data demands, computational effort, interpretability, and infrastructure.
- Demarcation: not a merely monetary concept; not an external secondary evaluation; not an absolute quantity independent of domain and enabling space.

Utility Function $U(M, D, t)$

- Short definition: a reconstructive evaluative structure that brings together approximate truth, explanatory power, and model costs within a domain.
- Function: serves to represent the domain-specific epistemic optimality or competitiveness of a model.
- Demarcation: not an algorithmic decision procedure; not a complete computation of scientific rationality; not a domain-independent ranking.

Relative Difference $\Delta U(M, M^*, D, t)^*$

- Short definition: a comparative quantity between a model M and an available alternative model M^* within the same relevant domain.
- Function: shows whether an alternative is only marginally better, distinctly superior, or stably epistemically dominant.
- Demarcation: not an independent theory; not an exact universal measure; not an evaluation without reference to $E(t)$.

Tolerance Range ϵ

- Short definition: a context-dependent range within which a model remains epistemically competitive despite a marginally lower utility value.
- Function: prevents minimal or unstable differences between models from being interpreted immediately as global falsification.
- Demarcation: not an arbitrary protective clause; not an immunization against criticism; not a fixed quantity independent of discipline and domain.

Relevant Domain

- Short definition: a current scientific area of use in which a model performs explanatory, predictive, interventionist, or structuring work.
- Function: distinguishes epistemically viable use of a model from merely historical, didactic, or illustrative residual use.
- Demarcation: not an arbitrary application context; not a merely cultural persistence; not a mere mention of a model.

Status and Domain of Validity

The refinements introduced in this paper do not constitute an independent extension of the Epistemics base canon in the strong sense. They serve exclusively as a local philosophy-of-science specification of central concepts for the special case of contextual and global falsification of scientific models. Their domain of validity is restricted to the model-evaluation architecture of this paper.

There is no silent extension, reinterpretation, or retroactive modification of the Epistemics base canon. The base canon remains unchanged in meaning, function, and demarcation.

Any future deviation, further refinement, or genuine canonical extension beyond this local conceptual specification must be explicitly indicated, locally bounded, and justified. Implicit shifts of meaning or informal canon extensions are excluded.

Bibliography

- Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press.
- Cartwright, Nancy. 1999. *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.
- Da Costa, Newton C. A., and Steven French. 2003. *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning*. Oxford: Oxford University Press.
- Duhem, Pierre. 1906. *La théorie physique: Son objet et sa structure*. Paris: Chevalier & Rivière.
- Einstein, Albert. 1905. "Zur Elektrodynamik bewegter Körper." *Annalen der Physik* 17: 891–921.
- Hacking, Ian. 1983. *Representing and Intervening*. Cambridge: Cambridge University Press.
- IPCC. 2021. *Climate Change 2021: The Physical Science Basis*. Cambridge: Cambridge University Press.
- Kepler, Johannes. 1609. *Astronomia Nova*. Heidelberg: Jonas Sabber.
- Kepler, Johannes. 1619. *Harmonices Mundi*. Linz: Johann Planck.
- Kitcher, Philip. 1981. "Explanatory Unification." *Philosophy of Science* 48 (4): 507–531.
- Kitcher, Philip. 1989. "Explanatory Unification and the Causal Structure of the World." In *Scientific Explanation*, edited by Philip Kitcher and Wesley C. Salmon, 410–505. Minneapolis: University of Minnesota Press.
- Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Lakatos, Imre. 1970. "Falsification and the Methodology of Scientific Research Programmes." In *Criticism and the Growth of Knowledge*, edited by Imre Lakatos and Alan Musgrave, 91–196. Cambridge: Cambridge University Press.
- Maxwell, James Clerk. 1865. "A Dynamical Theory of the Electromagnetic Field." *Philosophical Transactions of the Royal Society of London* 155: 459–512.
- Morgan, Mary S., and Margaret Morrison. 1999. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.
- Niiniluoto, Ilkka. 1987. *Truthlikeness*. Dordrecht: Reidel.
- Niiniluoto, Ilkka. 1998. "Approximation in Science." *Poznan Studies in the Philosophy of the Sciences and the Humanities* 63: 97–109.
- Oddie, Graham. 1986. *Likeness to Truth*. Dordrecht: Reidel.
- Oreskes, Naomi, Kristin Shrader-Frechette, and Kenneth Belitz. 1994. "Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences." *Science* 263 (5147): 641–646.
- Popper, Karl R. 1934. *Logik der Forschung*. Vienna: Springer.
- Popper, Karl R. [1934] 1959. *The Logic of Scientific Discovery*. London: Hutchinson.
- Psillos, Stathis. 1999. *Scientific Realism: How Science Tracks Truth*. London: Routledge.

- Quine, W. V. O. 1951. "Two Dogmas of Empiricism." *The Philosophical Review* 60 (1): 20–43.
- Rapp, Stefan. 2026a. *Epistemics: Model Management Under Finite Conditions*. Zenodo. <https://doi.org/10.5281/zenodo.18441326>.
- Rapp, Stefan. 2026b. *Friction: Boundary Signal of Finite Load-Bearing Capacity in Subjective, Intersubjective, and Functional-Empirical Stability Spaces*. Zenodo. <https://doi.org/10.5281/zenodo.18434699>.
- Rapp, Stefan. 2026c. *Revision under Finite Conditions: A Theory of Model Transformation in Epistemics*. Zenodo. <https://doi.org/10.5281/zenodo.18935928>.
- Rapp, Stefan. 2026d. *Efficient Search under Finite Conditions: A Dual-Mode Architecture of Model Management*. Zenodo. <https://doi.org/10.5281/zenodo.18799473>.
- Rapp, Stefan. 2026e. Domains, Boundaries, and Transition Functions: On the Domain-Relative Validity, Migration, and Coupling of Models under Finite Conditions. Zenodo. <https://doi.org/10.5281/zenodo.19542526>.
- van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford: Clarendon Press.
- Weisberg, Michael. 2013. *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.
- Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
- Worrall, John. 1989. "Structural Realism: The Best of Both Worlds?" *Dialectica* 43 (1–2): 99–124.

Appendix: Didactic Examples of the Two-Stage Falsification Structure

The following examples serve to illustrate didactically the distinction developed in this paper between contextual and global falsification. They are not a substitute for detailed historical analysis in particular fields but show by example how model validity, domain relevance, tolerance ranges, domain decomposition, and global model replacement can be reconstructed within the proposed framework. The examples are therefore not to be understood as complete historical reconstructions but as application sketches of the conceptual architecture of this paper.

A.1 Newtonian Mechanics: Contextual Falsification without Global Elimination

Newtonian mechanics is a paradigmatic case of contextual falsification. Its validity was not simply eliminated by relativistic physics and modern gravitational theory but rather restricted in a domain-specific way. At very high velocities, in strong gravitational fields, or in highly precise astronomical measurements, it loses epistemic competitiveness vis-à-vis relativistic models. In these domains, its approximate truth diminishes because its model structure no longer captures relevant system features with sufficient accuracy.

At the same time, Newtonian mechanics remains epistemically viable in many technical, everyday, and engineering domains. For bridge construction, mechanical engineering, many technical applications, and classical approximate calculations of celestial mechanics, it possesses high explanatory power, low model costs, and adequate predictive accuracy. Its didactic use in schools and universities is an additional matter, but does not on its own ground its epistemic competitiveness. In the language of this paper, this means: Newtonian mechanics is falsified or in need of revision in particular subdomains but remains competitive in other relevant domains.

The case shows why falsification must not be equated with global elimination. Relativistic physics does not replace Newtonian mechanics in every context of application; it constrains its domain of validity. The resulting structure is therefore not a simple refutation but a domain refinement: the model is preserved, but its validity is restricted.

Anti-immunization check. The preservation of Newtonian mechanics is not a mere protective strategy. Its remaining domains are stabilized independently through scientific and technical practice. Its error bounds are known, its costs low, and its conditions of application clearly specifiable. The domain restriction therefore generates new epistemic gains rather than artificially rescuing a weakened model.

A.2 Phlogiston Theory: Global Falsification with Historical Residual Use

Phlogiston theory serves as an example of global falsification. In the transition to modern chemistry, it lost performance not only in a narrowly bounded subdomain but was displaced in the relevant scientific contexts of explaining combustion, oxidation, and chemical reactions by more powerful alternatives. Its model costs rose, its explanatory power declined, and available alternatives offered stably better connections.

In the proposed framework, this means: phlogiston theory did not remain epistemically competitive in any relevant scientific domain. Its present-day use is historical, didactic, or related to the history of science, but no longer part of current chemical model evaluation. It can still be explained, taught, or used as an example of scientific change, but this residual use does not ground continuing epistemic competitiveness.

The case shows the distinction between historical persistence and a relevant domain. A model can remain culturally, didactically, or historically present and yet be globally falsified. Global falsification therefore does not mean that a model disappears from all discourses but that it is no longer viable in any current scientific domain against available alternatives.

Anti-immunization check. An artificial residual domain such as "phlogiston as a historical figure of thought" would not ground epistemic competitiveness. It is not a current chemical application domain but a didactic or historical form of use. This is precisely what shows why relevant domains must not be derived from mere residual use.

A.3 Climate Models: Contextual Falsification within Model Families

Climate models show how contextual falsification operates in complex model families. A climate model or an ensemble of climate models can possess high viability in particular areas, such as long-term global temperature trends, while other subdomains exhibit greater uncertainties—for instance, regional precipitation patterns, extreme events, or short-term climate variability. Deviations in such a subdomain do not necessarily falsify the entire model family but can affect particular parameterizations, submodels, or weightings.

Within the framework of this paper, this means that falsification often operates redistributively. It shifts epistemic weights within a model portfolio, weakening some components and strengthening others. An incorrect regional prediction can lower the approximate truth of a submodel in a specific domain without eliminating the entire model architecture globally. The consequence is then local revision, reweighting, or domain refinement.

This case shows particularly clearly why modern modeling practice cannot be captured by a simple Popperian elimination logic. In simulation-based sciences, models are often modular, parameterized, and ensemble-organized. Falsification then rarely affects an overall model as a unified structure but operates within a structured model space.

Anti-immunization check. The preservation of a climate model or ensemble is rationally reconstructible only when the remaining domains are independently supported by data, testing procedures, and contexts of application. Merely shifting problematic predictions into ever smaller subdomains would not be sufficient. Legitimacy arises only when the domain refinement generates new diagnostic, predictive, or explanatory gains.

A.4 Ptolemaic Epicycles: A Boundary Case between Adjustment and Ad Hoc Stabilization

Ptolemaic epicycle models are well suited as a boundary case because they show both empirical adaptability and possible ad hoc stabilization. Such models could describe observed planetary motions over long periods through additional geometric constructions. In particular historical enabling spaces, they were therefore not simply irrational or epistemically empty. Their stability rested on mathematical adaptability, institutional embedding, and the absence of competing, stable alternatives with higher overall performance.

With the development of Keplerian and later Newtonian models, however, the epistemic enabling space shifted. New mathematical, empirical, and theoretical conditions made alternatives available that not only reproduced observations but enabled a deeper structural and dynamic explanation. As a result, the Ptolemaic model class lost its relative competitiveness in the relevant astronomical and physical domains.

This case shows the importance of the anti-immunization logic. Not every model adjustment is illegitimate. An adjustment is epistemically productive when it makes new testable distinctions visible, separates sources of error, or increases explanatory power. It becomes problematic when it merely absorbs known counter-evidence without generating new gains. The boundary between legitimate model refinement and ad hoc rescue thus does not lie in the mere fact of additional assumptions but in their epistemic productivity.

Anti-immunization check. Epicycles become problematic when additional constructions accomplish only defensive adjustments and no longer open up new explanatory or predictive structures. Their global replacement does not arise from a single observation but from stable relative inferiority to alternatives that, under changed conditions of $E(t)$, became available and more viable.

A.5 The Four Humors Theory: Global Replacement despite Cultural Persistence

The four humors theory exemplifies how a model can persist culturally without remaining scientifically competitive. Historically, it offered a comprehensive ordering of medical observations, interpretations of disease, and therapeutic practices. Within its enabling space at the time, it could function as a stabilizing framework, since alternative physiological, microbiological, or biochemical models were not yet available in their current form.

Under modern conditions, however, the four humors theory possesses no epistemic competitiveness in medical diagnostics, therapy, or disease modeling. Its explanatory and interventionist performance has been replaced by modern medicine, physiology, pathology, microbiology, and evidence-based procedures. Present-day appeals to the four humors theory may occur in historical, cultural-historical, or alternative-medicine contexts, but they do not ground a relevant scientific domain in the sense of this paper.

The case again illustrates the difference between persistence and validity. Models do not necessarily disappear from language, culture, or history when they are scientifically falsified. Global falsification means rather that a model no longer possesses viable competitiveness within the relevant scientific contexts of practice.

Anti-immunization check. The four humors theory cannot be epistemically rescued by treating its present-day historical or metaphorical use as a relevant medical domain. Such a residual domain would not be independently stabilized by current medical testing procedures, therapeutic successes, or scientific evaluative orderings.

A.6 Overview of the Examples

Case	Problematic domain	Preserved domain	Diagnosis within the framework of this paper
Newtonian mechanics	relativistic velocities, strong gravitation, high-precision boundary cases	technical, everyday, and engineering applications, and classical regimes of approximation	contextual falsification with domain restriction
Phlogiston theory	combustion, oxidation, chemical reaction theory	no current scientific domain; historical and didactic residual use	global falsification
Climate models	particular regional, short-term, or highly uncertain subdomains	long-term global trends and robust ensemble-based projections	contextual falsification within model families
Ptolemaic epicycles	astronomical explanation under changed mathematical and physical conditions	historical and didactic reconstruction	global replacement; boundary case between legitimate adjustment and ad hoc stabilization
Four humors theory	modern medical diagnostics and therapy	historical, cultural-historical, or metaphorical use	global falsification despite cultural persistence

The examples show that the two-stage falsification structure is not merely a formal distinction. It allows different cases of scientific model evaluation to be reconstructed in differentiated ways: models can fail locally and yet be rationally preserved; they can persist historically and yet be globally falsified; they can be improved through legitimate domain refinement or stabilized artificially through merely ad hoc decomposition. What is decisive in each case is whether the remaining model validity is supported by relevant domains, disciplinary evaluative standards, and new epistemic gains.