Computationally Reframing the Theory-Ladenness of Observation

Hanson's Theory-Ladenness, Predictive Processing, and the Bayesian Structure of Scientific Discovery

Mako Yamaguchi* & Hajime Sugio†

December 21, 2025

Abstract

This paper offers a computationally explicit reformulation of N. R. Hanson's thesis that observation is theory-laden *seeing-as* [16]. Drawing on predictive processing (PP) and Bayesian modeling [3, 11, 20], it advances a *failure-sensitive bridge hypothesis*: a role- and constraint-preserving correspondence—not a reductive identification—between personal-level philosophical roles and subpersonal computational roles.

Crucially, the paper rejects the slogan "theory = prior." Instead, it models theories as *meta-level admissibility constraints* on model space: they constrain eligible variables, admissible structures, likelihood/measurement assumptions, and relatively durable precision expectations. On this view, what can count as an observation report or a scientific "fact" is shaped by inference and policy-governed reporting under theory-conditioned constraints together with stabilized measurement and reporting regimes.

The first part reconstructs Hanson's analysis in computational terms: *seeing-as* is modeled as posterior-guided inference under a theory-conditioned generative model, and observation statements are treated as goal-, context-, and norm-sensitive *report policies* over posterior states rather than as direct readouts of raw data. The second part recasts discovery dynamics as multi-tier revision: apparent mismatches may be resolved by *within-model* routes (parameter retuning, measurement/noise-model repair, and precision reallocation), but persistent *structured* residuals—stable across plausible measurement/noise variants—can rationally motivate *between-model* selection and, when required, revision of the admissible model space itself. A reanalysis of Kepler's adoption of elliptical orbits illustrates how such constraint changes reorganize the space of admissible observations, facts, and explanations.

Overall, the paper proposes a computationally explicit yet non-reductive framework that clarifies structural relations among theory, perception, and discovery.

Keywords: Theory-ladenness of observation; Seeing-as; Predictive processing; Bayesian modeling; Hierarchical generative models; Precision weighting; Admissibility constraints (model space); Report policies; Model evidence and model comparison; Abduction (IBE); Scientific discovery and theory change.

*E-mail: 43mako@gmail.com
†E-mail: hajime-sugio@sophia.ac.jp; Department of Philosophy, Sophia University, Tokyo, Japan

1 Introduction

1.1 Background of the Present Study: Is Observation Neutral?

Where, ultimately, does the warrant for scientific knowledge lie? Since the rise of modern science, one influential answer has been *observation*. A familiar *idealized* picture—often invoked in textbook reconstructions of empiricism and logical empiricism—treats observation reports as providing a comparatively neutral basis from which laws can be inductively inferred and theories can be tested. On this picture, the observer is sometimes portrayed as approximating a *tabula rasa*: as if a camera could passively register "bare facts" from sensory stimulation prior to any conceptual articulation. (We emphasize that this is a deliberately simplified foil, not a precise exegetical claim about every empiricist position.)

However, it is doubtful that this "neutral observer" model is compatible with either actual scientific practice or ordinary perceptual experience. Even when the proximal stimulus is held fixed, what is *seen* can vary with the observer's conceptual repertoire, background expectations, and practical interests. Classic ambiguous figures make this vivid: the very same drawing can be seen as a young woman or as an old woman, depending on how it is organized perceptually [19]. Philosophically, this motivates the familiar worry that theory-testing by appeal to observational "facts" risks circularity: if the relevant background commitments partly determine what counts as an observational fact, then the evidence seems to presuppose what it is meant to test [16, 24, 32].

At the same time, the moral need not be that evidence is hopelessly subjective. Scientific practice routinely distinguishes messy, theory-mediated *data* from comparatively stable *phenomena* extracted from those data across methods, instruments, and laboratories [2]. This suggests that the relevant question is not whether observation is *influenced* by background commitments (it often is), but *how* such influences operate and under what conditions they threaten (or fail to threaten) objectivity.

Contemporary cognitive science likewise emphasizes that perception is not a passive intake of sensory inputs, but an inferential and constructive process. Gregory's well-known slogan that "perceptions are hypotheses" captures the idea that perception involves active inference toward a best interpretation of sensory evidence [15]. This general orientation has been developed in explicitly Bayesian and predictive frameworks, in which perceptual content is understood as inference to latent causes under a generative model [3, 10, 23, 25, 31]. These developments resonate with philosophical doubts about the neutrality of observation and motivate a re-examination of what "observation" amounts to in scientific inquiry.

The present study addresses these issues under the heading of the *theory-ladenness of observation*. The guiding question is whether observation is not merely an optical and physiological process, but an active inferential process mediated by conceptual and cognitive frameworks. This question is at once a challenge to how scientific objectivity is to be understood and a methodological problem concerning the appropriate level of description (*personal* vs. *sub-personal*) at which observation should be theorized [5, 27].

Preview of the proposal. The paper advances a *failure-sensitive bridge hypothesis* between personal-level roles (theory, seeing-as, observation statements, anomaly, and discovery) and sub-personal computational roles (hierarchical generative inference, precision-weighted mismatch, and model updating/selection). We reject the slogan *theory* = *prior* by treating theories as *meta-level admissibility constraints* on model space (Sections 1.5 and 1.7). We also treat observation statements as *report policies* that extract publicly assessable commitments from posterior states under goals, contexts, and community constraints (Section 4.3;

see also Section 1.6), rather than as direct readouts of theory-free data.

Contributions in brief (what this paper adds). The contribution of the paper is not a general endorsement of PP/FEP, but a set of computationally explicit *bridge claims* that go beyond a change of vocabulary. First, it replaces the slogan "theory = prior" with a *typed decomposition* of theory-ladenness across (i) theory-level *admissibility constraints* on model space ($T = \langle \mathcal{V}_T, \mathcal{S}_T, \mathcal{L}_T, \Pi_T \rangle$), (ii) concrete instrumentation/calibration regimes (I), and (iii) community-level reporting norms (N), with explicit boundary rules separating these components (Sections 1.5 and 1.6; cf. Table 1). Second, it models observation statements as *report policies* over posterior states, making the public, norm-governed character of scientific reporting explicit without requiring a strong thesis of cognitive penetration (Section 4.3). Third, it operationalizes the anomaly–discovery connection via a multi-tier revision picture: routine misfit is handled by within-model routes, whereas *persistent structured residuals* motivate between-model selection and, when required, admissibility-constraint revision (Sections 1.4 and 1.12; see also Section 5.2). Because the bridge is stated with success conditions, failure modes, and discriminators (C1–C3; F1–F2; D1–D3), it yields testable and failure-sensitive predictions rather than a mere analogy.

Target of theory-ladenness (three layers). To reduce ambiguity about what is meant by "theory-ladenness," we distinguish three layers at which it can arise.

1. **Early input-processing (hard cases).** When low-level sensory/instrumental channels are assigned high precision, perception is strongly constrained by input, and theory-ladenness is expected to be limited in scope (cf. D1).

2. **Mid–high-level generative inference (seeing-as as posterior organization).** Theory-ladenness is expected to be salient at levels where categorical organization and higher-level hypotheses structure posterior inference under a theory-conditioned model space (Sections 1.5 and 3).

3. **Report and public commitment (policy- and norm-governed).** Theory-ladenness is often most institutionalized at the level of observation statements, where reporting thresholds, hedging, and admissible vocabulary are regulated by goals, contexts, and community constraints (Section 4.3; cf. D3).

The present paper focuses primarily on (2) and (3), while treating (1) in a graded way via precision allocation (D1). Accordingly, the bridge does *not* presuppose a strong thesis of cognitive penetration of early vision; it is compatible with skeptical assessments that locate many "top-down" effects at attentional, inferential, or reporting stages [9].

1.2 The Evolution of the Concept of Observation in Philosophy of Science

Early twentieth-century philosophy of science often attempted to separate "observational" from "theoretical" vocabularies and to assign observation reports a distinctive epistemic role. The post-positivist turn challenged this separation: Sellars rejected the Myth of the Given, Kuhn emphasized paradigm-dependent perception and description, and Feyerabend denied the autonomy of an observation language [8, 24, 32]. Hanson—the focal point for the present study—crystallized the issue by analyzing observation as "patterned seeing" and by linking seeing-as to discovery dynamics [16].

1.3 Hanson's Challenge: Theory-Ladenness and the Philosophy of Discovery

Hanson's core claim is that "seeing" and "seeing-as" are inseparable: the content of observation is structured from the outset. His canonical historical illustration contrasts the ways in which Tycho Brahe and Kepler, confronted with the same astronomical data, "saw" different celestial motions—not because they received different stimuli, but because they organized those stimuli under different theoretical commitments [16, 38]. The intended lesson is not that observers fabricate facts at will, but that observation is already structured by prior conceptual organization.

Drawing on Gestalt psychology, Hanson emphasized that shifts in organization can be both sudden and systematic, as in classic cases of perceptual reorganization [19]. Theory-ladenness, on this view, is not merely a source of bias but a positive force that can drive scientific discovery. As Peircean abduction (and later discussions of inference to the best explanation) illustrate, novel hypotheses are often generated through concept-laden engagement with anomalous patterns in what is observed [17, 30].

Nonetheless, Hanson's account—rich in epistemological insight—does not by itself explain the cognitive or computational mechanisms that realize such "seeing-as." In particular, the question of how "theory" might intervene in perception opens a gap between descriptions at the *personal* level (scientists' reasoning, judgment, and discourse) and descriptions at the *sub-personal* level (neural and computational information-processing mechanisms) [5, 27].

The present study does not attempt to collapse these levels into a reductive unity. Rather, it proposes a *bridge hypothesis*: a *structural correspondence* between explanatory roles at the personal level and computational roles at the sub-personal level. Here "structural correspondence" is understood in a modest sense—as a role- and constraint-preserving mapping rather than a strict mathematical isomorphism. That is, the question is what kind of correspondence can be identified between Hanson's observation–anomaly–discovery pattern and the predictive-processing cycle of generative inference–(precision-weighted) prediction error–belief updating and model comparison [1, 10, 11, 26, 31].

The correspondence envisioned here is analogous to Marr's distinction among computational, algorithmic, and implementational levels [27], Dennett's relations among the intentional, design, and physical stances [7], and Craver's mechanistic account of multi-level explanation [5, 22]. Descriptions at the personal level concern the normative and semantic structure of how scientists observe, form hypotheses, and revise theories. Descriptions at the sub-personal level specify neurocomputational constraints that make such practices possible. To "bridge" these levels is not to assert identity, but to clarify which structural relations must obtain for the two levels of description to *mutually constrain* one another.

From this perspective, the proposed correspondence is motivated by at least three shared features: (1) both levels treat the tension between input and expectation as a primary driver; (2) surprise or anomaly functions as a trigger for structural revision; and (3) the dynamics of revision plausibly involve both continuous learning and, in some cases, discrete shifts in model class (model selection) [11, 24, 26].

Finally, clarifying how "theory" intervenes in perception requires a careful reconsideration of the relation between theories and priors. As developed in Sections 1.5 and 1.7, we treat theories as meta-level structures that constrain the space of admissible generative models—fixing representational vocabulary, admissible causal structure, and likelihood/measurement assumptions—rather than adopting the slogan *theory* = *prior* [2, 12, 26, 29, 40].

1.4 Bridge Hypotheses and Criteria for Structural Correspondence

Because structural correspondence can otherwise be mistaken for a mere analogy, this paper makes explicit what would count as a *successful* bridge between the Hanson-style personal-level pattern and the PP-style sub-personal computational pattern. In what follows, the correspondence is treated as *constraint-bearing*: it should not only "match" roles, but also articulate when and why the match fails [5, 22].

Type of correspondence. The correspondence proposed here is not a strict identity claim. It is a role-to-role mapping intended to preserve key explanatory constraints. In particular, it aims to preserve: (i) what drives revision (anomaly/surprise), (ii) what kinds of revision are available (within-model vs. between-model), and (iii) what changes when revision is radical (the space of admissible facts and explanations).

Success conditions (minimal criteria). The bridge hypothesis is intended to succeed only if the following conditions are met:

C1 (Functional drive) A mismatch between expectation and input plays a driving role in both descriptions: anomalies at the personal level correspond to sustained, precision-weighted prediction-error regimes that motivate revision at the sub-personal level [10, 11].

C2 (Two-tier dynamics) The account distinguishes *within-model* updating (adjusting parameters or priors within a fixed model class) from *between-model* selection (shifting to a different model class when the existing class cannot accommodate the relevant discrepancies) [1, 26].

C3 (Reorganization) Between-model shifts reorganize what can count as an admissible observation, fact, or causal explanation—i.e., they restructure the relevant explanatory space rather than merely refining a fixed description [16, 24].

Failure modes (how the correspondence can break). Conversely, the correspondence is *not* intended to hold in at least the following kinds of cases:

F1 The Hanson-style transition from anomaly to discovery is present at the personal level, yet no change of model class is needed at the computational level (e.g., the discrepancy is resolved entirely by attentional reweighting or post-perceptual judgment without requiring model-space revision).

F2 A computational model-selection episode occurs, yet there is no corresponding shift in seeing-as (no change in the relevant semantic or explanatory framework), so that the episode amounts only to parameter identification rather than a reorganization of interpretation.

Two immediate constraints. Stating these criteria yields two immediate constraints that guide the remainder of the paper. First, theory-ladenness does not entail unconstrained top-down influence on early vision; within predictive-processing formulations, the extent of top-down influence depends on hierarchical level and on the allocation of precision (the estimated reliability of prediction errors) [3, 10, 20]. This constraint helps keep the present proposal compatible with skeptical assessments of strong "cognitive penetration" claims [9]. Second, "theory" must be treated as a meta-constraint on model space rather than as a numerical prior, to avoid a category mistake (see Section 1.7). Later sections will operationalize

these constraints in more detail—including a minimal account of observation reports as *report policies* that extract actionable public commitments from posterior states under goals (G), contexts (C), and community-level constraints, rather than as a direct probability-to-text mapping [4, 11].

Public scientific theories vs. internal generative models (no category mistake). A further worry is that scientific theories are public symbolic and institutional objects (mathematics, representations, instruments, and shared standards), whereas PP-style generative models are posited as sub-personal resources in individual agents. The bridge hypothesis does *not* identify these as the same kind of entity. Rather, it targets a role- and constraint-preserving correspondence between (i) the public, norm-governed theoretical *package* that structures scientific practice and (ii) the internal inferential resources that must be in place for individual scientists to competently perceive, model, and report under that package. In the present notation, T is an *as-if* representation of those practice-stabilized constraints as they become operative for an agent (typically via training and enculturation), while N and I keep the communal reporting norms and instrumentation/calibration regimes explicitly in view rather than reducing them to neural states. Accordingly, the bridge concerns how public theory-guided practices can *constrain* and be *realized by* sub-personal inference without collapsing levels or treating theories as literal probability distributions [5, 7, 22, 27].

Mediation hypothesis (how public theory constrains individual inference). The bridge presupposes a minimal mediation story: public, norm-governed theory packages (T, N, I) constrain individual sub-personal inference *via training and practice*. Roughly: (i) enculturation and disciplinary training fix an agent's usable representational vocabulary and structural expectations (linking to $\mathcal{V}_T$ and $\mathcal{S}_T$), (ii) instrument pipelines and calibration routines stabilize how licensed measurement families are instantiated and which error channels are treated as diagnostically reliable (linking to I, $\mathcal{L}_T$, and durable Π_T; cf. Section 1.6), and (iii) reporting norms shape the learned space of permissible report options, thresholds, and vocabulary in the report-policy transition from posterior states to public commitments (linking to N and Section 4.3). On this view, the correspondence is neither a mere metaphor nor an identity claim: it is a role-preserving mapping that becomes operative *only under such mediated stabilization*; when the mediation is absent or disrupted (e.g., novices, cross-community mismatches, unstable calibration), the bridge can fail in the sense captured by the stated failure modes.

1.5 A Minimal Specification: Theory as Meta-Level Constraints on Model Space

To reduce the risk that "theory as a meta-constraint" remains a slogan, we provide a minimal specification. Let $\mathcal{M}$ denote a space of candidate generative model classes M. A *theory* T is not identified with a single probability distribution; rather, it is modeled as a constraint operator that selects an admissible subset:

$$\mathcal{M}_T \subseteq \mathcal{M}.$$

A natural objection is that such restrictions are equivalent to imposing a degenerate $0/1$ model prior; we address this explicitly in Section 1.7. This formulation mirrors a familiar Bayesian distinction between (intra-model) parameter learning and (inter-model) choice of model class, now reinterpreted as a locus for theory-level constraint [1, 12, 26]. Concretely, we decompose the constraint into four components:

$$T \equiv \langle \mathcal{V}_T,\ \mathcal{S}_T,\ \mathcal{L}_T,\ \Pi_T \rangle,$$

where:

1. **Vocabulary / variable constraints ($\mathcal{V}_T$).** Which latent variables are eligible as candidate causes (what can be represented as a cause at all) [29, 40].

2. **Structural constraints ($\mathcal{S}_T$).** Which dependency/causal architectures are admissible (e.g., which links in a hierarchical generative structure are permitted) [29, 40].

3. **Likelihood / measurement constraints ($\mathcal{L}_T$).** Which likelihood families or noise/measurement models are treated as appropriate (how measurements are interpreted as data) [2, 12].

4. **Structural precision expectations (Π_T).** Relatively durable, methodologically stabilized expectations about which channels/levels are treated as trustworthy within the admissible model space (e.g., which discrepancies are taken as diagnostically significant) [3, 10]. Moment-to-moment, task- and goal-driven precision control is treated separately under goals (G) and contexts (C) later in the paper.

The admissible model space can then be represented schematically as

$$\mathcal{M}_T \;=\; \{\, M \in \mathcal{M} \;:\; \mathcal{V}(M) \subseteq \mathcal{V}_T,\; \mathcal{S}(M) \in \mathcal{S}_T,\; \mathcal{L}(M) \in \mathcal{L}_T,\; \Pi(M) \approx \Pi_T \,\}.$$

Types of theory change (constraint change). On this view, "theory change" can occur by: (i) expanding/revising $\mathcal{V}_T$ (new kinds of causes become representable), (ii) relaxing or replacing $\mathcal{S}_T$ (new structures become admissible), (iii) revising $\mathcal{L}_T$ (new measurement/noise interpretations), or (iv) redistributing Π_T (what counts as reliable evidence shifts). This decomposition lets us state more precisely what is meant by "theory reorganizes observation and facthood": it reorganizes the admissible model space $\mathcal{M}_T$ in which posterior inference and reporting take place.

What is (and is not) counted as "theory" in T. In this paper, $T = \langle \mathcal{V}_T, \mathcal{S}_T, \mathcal{L}_T, \Pi_T \rangle$ is a *narrow* notion of theory: it specifies constraints on admissible representational resources (variables), dependency/causal architecture, and measurement/noise interpretation, plus durable precision expectations. This specification intentionally does *not* absorb all normative and institutional aspects of scientific practice into T. In particular, community-level reporting norms and instrumentation/calibration regimes are treated separately (later as N and I), entering explicitly in the policy-governed transition from posterior states to public observation statements (Section 4.3) [2].

A broader Kuhnian notion of a *paradigm* plausibly includes not only T but also such communal constraints [24]. When that broader reading is intended, one may schematically treat a paradigm as (T, N, I): a package of admissibility constraints together with stabilized measurement practices and reporting norms.

1.6 Boundary Rules for T, N, I, G, C

To keep the bridge model from becoming an unconstrained "everything-box," we fix explicit boundary rules for the components that enter the mapping.

(1) T (theory as admissibility constraints). $T = \langle \mathcal{V}_T, \mathcal{S}_T, \mathcal{L}_T, \Pi_T \rangle$ specifies what is *admissible in principle* within a community or research tradition: eligible variables ($\mathcal{V}_T$), admissible dependency architectures ($\mathcal{S}_T$), licensed likelihood/measurement families ($\mathcal{L}_T$), and relatively durable precision expectations (Π_T). Importantly, $\mathcal{L}_T$ is a *class-level license* (a "permit list"), not an instrument-specific calibration.

(2) I (instrumentation/calibration as instances). I denotes the *instance-level* measurement and calibration regime: a concrete instrument pipeline, preprocessing conventions, calibration history, and parameterization choices that instantiate (and sometimes stress-test) a licensed likelihood family. A simple way to see the distinction is:

$$\mathcal{L}_T \text{ licenses a family } \{p(s \mid z, M, \lambda)\}_{\lambda \in \Lambda}, \qquad I \text{ fixes or constrains } \lambda \text{ (and data handling) in practice.}$$

Thus, changes in $\mathcal{L}_T$ expand/alter what measurement models are licensed *in principle*, whereas changes in I revise how an accepted measurement family is concretely instantiated and validated.

(3) N (community-level reporting norms). N denotes public, socially enforced norms that regulate what counts as an acceptable observational claim: admissible vocabulary, hedging conventions, confidence thresholds, error-reporting standards, and "what must be reported" requirements. These norms constrain not only which report options are permitted, but also how public commitments are scored (e.g., penalties for overclaiming vs. underreporting).

(4) G, C (local goals and contexts). G and C are *local* and can vary within the same community: task goals (archiving, warning, hypothesis testing), situational contexts, time pressure, and pragmatic stakes. They modulate reporting behavior even when T, N are held fixed.

(5) Π_T vs. task-driven precision control. Π_T captures relatively durable, methodologically stabilized expectations about which channels/levels are treated as trustworthy. Moment-to-moment precision reallocation driven by attention, task demands, or context is treated under G, C (and is separated from Π_T) in order to keep the model components non-overlapping.

Locating components of theory-ladenness. Table 1 summarizes how the bridge framework locates theory-ladenness across admissibility constraints (T), instrumentation and calibration regimes (I), and community-level reporting norms (N), drawing on the data/phenomena distinction and standard Bayesian modeling practice (including model checking and comparison) [2, 12, 26].

Mini-example: detection claims as a joint function of T, I, and N. Suppose an instrument delivers a time series s (a noisy waveform). A minimal generative setup posits a latent cause $z \in \{\text{signal}, \text{noise}\}$ with a measurement model $p(s \mid z, M)$ and a prior $p(z \mid M)$. Theory-level constraints T determine (i) whether "signal" is even an admissible latent variable ($\mathcal{V}_T$), (ii) which families of noise/likelihood models are permitted ($\mathcal{L}_T$), and (iii) how precision is allocated to different channels (Π_T). Instrumentation and calibration regimes I fix concrete parameterizations of the measurement model (e.g., baseline corrections), while reporting norms N fix decision thresholds for public claims (e.g., when to report a detection vs. an upper bound) [2, 12].

Table 1: Locating components of theory-ladenness in the bridge framework.

Target phenomenon	**Primary constraints**	**Computational locus**	**Empirical/practical handle**
Seeing-as (perceptual/content level)	T (esp. $\mathcal{V}_T, \mathcal{S}_T, \Pi_T$)	Posterior structure $p(z \mid s, M)$ under a theory-conditioned model space	Precision manipulations; ambiguity vs. high-SNR regimes (D1)
Observation statement (public report)	G, C plus N, I	Report policy over posterior states (Section 4.3)	Thresholds/hedges; vocabulary norms; detection conventions (D3)
Scientific "fact" (stabilized report class)	N, I (plus shared T)	Reproducible posterior-supported reports across agents/labs	Calibration protocols; replication; error bars; inter-lab robustness
Anomaly	T and I (what counts as misfit)	Persistent, structured residuals after best within-model tuning (D2)	Posterior predictive checks; residual diagnostics; robustness to noise-model variants
Discovery / theory change	Change in T (and sometimes N, I)	Between-model selection / model-class revision	Model comparison under revised admissibility constraints; reorganized report space

Boundary note. $\mathcal{L}_T$ is a class-level license of admissible likelihood/measurement families, whereas I is an instance-level calibration and instrument pipeline that instantiates (and can stress-test) such families. N encodes community reporting norms; G, C encode local goals and contexts. See Section 1.6.

In such cases, the same raw readout s can license different public observation statements without any appeal to "theory-free data": changing I (calibration) or $\mathcal{L}_T$ (noise model) can change whether residual structure counts as systematic or as noise; changing N (thresholds) can change whether one issues a categorical claim, a hedged claim, or a non-detection. This illustrates concretely how the bridge locates theory-ladenness across model-space constraints (T), measurement practices (I), and public reporting norms (N), rather than collapsing it into "theory = prior."

1.7 Why Admissibility Constraints Are Not (Just) Priors

A natural Bayesian objection is that restricting an admissible model space $\mathcal{M}_T \subseteq \mathcal{M}$ is equivalent to imposing a degenerate model prior: simply assign $p(M \mid T) = 0$ for $M \notin \mathcal{M}_T$ and renormalize. Formally, one may define

$$p(M \mid T) \;\propto\; \mathbf{1}[M \in \mathcal{M}_T]\, \tilde{p}(M),$$

so that "admissibility" appears as a $0/1$ filter on a baseline model prior $\tilde{p}(M)$. Likewise, one may represent theory-level constraints via hierarchical constructions (hyperpriors) or structural priors that generate families of models, e.g. $p(M) = \int p(M \mid \eta)\, p(\eta)\, d\eta$, thereby embedding model-space restrictions within standard Bayesian machinery [12, 26].

The key point: formal embeddability does not fix philosophical role. We grant the formal point: *constraints can be encoded probabilistically*. However, the present thesis is not an identity claim of the form *theory = (model) prior*. It is a *role claim* about what theories do in scientific practice. Priors (whether

over parameters or over models) presuppose that the relevant representational vocabulary, admissible structures, and measurement interpretations are already in place. By contrast, what is called "theory" here is precisely the package that helps *establish and stabilize* those prerequisites: it constrains what counts as an eligible variable, what dependency architectures are admissible, which likelihood/measurement families are licensed, and which precision expectations are methodologically treated as durable.

Role-identification: constitutive vs. regulative. The issue is not merely that constraints can be *encoded* as priors, but that they play a different *identifiable role* in scientific practice. Priors are *regulative*: they weight options *within* an already articulated inferential setting—where the representational vocabulary, admissible structures, and measurement interpretations are treated as fixed for the purposes of updating. By contrast, admissibility constraints are *constitutive*: they help *fix* what counts as an eligible variable, an admissible dependency architecture, and a licensed likelihood/measurement family in the first place (cf. $\mathcal{V}_T, \mathcal{S}_T, \mathcal{L}_T$ and the T/I boundary in Section 1.6). A practical diagnostic is this: if changing the relevant assumption changes what is even *well-typed* as a hypothesis or as a measurement interpretation, then it is constraint-like (constitutive), not merely a reweighting within a fixed space.

Revision target: why this matters for C2/C3. The constitutive/regulative contrast matters because the bridge criteria explicitly require a difference in *targets of revision*. Within-model and within-space adjustments preserve the underlying vocabulary/structure/measurement interpretation and yield continuous re-tuning (criterion **C2** in Section 1.4). Constraint revision targets the space itself—changing what is admissible as a hypothesis and thereby reorganizing what can count as an observation report, a stabilized fact, or an acceptable explanation (criterion **C3**). This is precisely what the slogan *theory = prior* tends to obscure: it makes genuine reorganization look like nothing more than reweighting within a fixed space.

Optional: graded admissibility (soft licenses). In practice, admissibility need not be a strict $0/1$ filter. One can model this by a licensing strength $\alpha_T(M) \in [0, 1]$ that expresses methodological permissibility within a research tradition, so that highly licensed model families are treated as more eligible than marginal ones. Crucially, α_T is not introduced as a private degree of belief; it tracks *publicly stabilized licenses* (what is methodologically allowed) and is revised by methodological change, not by ordinary within-space Bayesian updating.

As-if representation and the bridge target. For bridge purposes, it is often convenient to represent a theory T as if it were a filter or generator of admissible models, and (when needed) to encode that filter as a structural prior or a hyperprior. But doing so is not to collapse semantic and methodological commitments into a mere probability assignment. Rather, it provides an *as-if computational representation* of the constraint-bearing roles that theories play at the personal level. This is why the slogan "theory = prior" is rejected as a category mistake: the mistake is to identify a public, norm-governed theoretical package (including representational, methodological, and measurement licenses) with a literal probability distribution (a prior) in an individual agent. Accordingly, the bridge does not claim that theories *are* probability distributions, but that theories can be modeled as constraints that *shape* the space of admissible models (hence the families of priors and likelihoods) on which Bayesian updating and model comparison operate.

Consequence for theory change. On this reading, "changing the theory" is not merely shifting weights within a fixed space (reweighting priors over already admissible hypotheses). It can also involve revising the admissibility package itself $T \mapsto T'$ (hence $\mathcal{M}_T \mapsto \mathcal{M}_{T'}$), thereby reorganizing what counts as an admissible explanation and a reportable observation. This is the sense in which constraint revision differs from (and can rationally be motivated beyond) ordinary within-space updating, even though both can be expressed in Bayesian notation [12, 26].

1.8 The Predictive Turn in Contemporary Neuroscience

In recent decades, computational neuroscience has developed a family of Bayesian frameworks—including predictive coding, predictive processing (PP), and free-energy/active-inference formalisms—that provide candidate sub-personal mechanisms for the kind of "seeing-as" emphasized in the theory-ladenness tradition [3, 10, 11, 20, 31]. According to these approaches, the brain (or cognitive system) maintains hierarchical generative models of the latent causes of sensory inputs and reduces mismatches between predicted and incoming signals by updating beliefs, often described in terms of (precision-weighted) prediction errors [10, 11].

A brief terminological note will reduce confusion. In this paper, *predictive coding* refers to a family of mechanistic architectures (often discussed in relation to cortical hierarchies); *predictive processing* refers to the broader research program that interprets perception and action in predictive terms; *active inference* extends the story to action and policy selection within the free-energy formalism; and the *free-energy principle* (FEP) provides a general variational framework often used to unify these ideas [4, 10, 11]. For expository purposes, we will sometimes write prediction error in the schematic form "actual minus predicted." This is a *didactic idealization*: in standard formulations, error signals are typically precision-weighted and defined across multiple hierarchical levels [10, 11].

Within PP, perception is construed as estimation of causes via a posterior distribution, and high-level priors (together with precision allocation) can exert strong constraints on perceptual content [3, 10, 23, 25]. This provides one route to Seth's "controlled hallucination" metaphor, while also motivating clinical applications that interpret certain psychopathologies as disorders of prediction and precision [34, 35]. Importantly, none of this entails that cognition exerts unconstrained influence on early perceptual processing; the relevant claim is conditional and hierarchy-/precision-sensitive, and the empirical debate about "cognitive penetration" remains live [9].

1.9 Research Question: Can Hanson's Theory-Ladenness Be Computationally Reconstructed?

Against this background, the central question of the present study may be stated as follows:

> **Can Hanson's theory-ladenness of observation and the process of scientific discovery be computationally reconstructed within the predictive-processing framework, while respecting the distinction between the personal and sub-personal levels?**

The aim of this study is to integrate Hanson's philosophical analysis with the computational architecture of PP, reinterpreting observation, hypothesis formation, and theory change in terms of Bayesian inference and model updating and selection. In doing so, the paper develops and evaluates a bridge hypothesis: it

clarifies the structural correspondences—in the sense made explicit in Section 1.4—between personal-level patterns of scientific reasoning and discovery and the sub-personal inferential machinery posited by predictive processing. The result is intended to be a naturalized yet non-reductive model: it seeks level-coherent constraints between philosophy of science and computational cognitive science without collapsing one level into the other [5, 22, 27].

1.10 Relation to Bayesian Philosophy of Science and the Paper's Novelty

Bayesian philosophy of science supplies formal tools for within-model updating and between-model comparison. Our project is continuous with that tradition at the level of formal structure, but it differs in target and level-relations: we use Bayesian/PP roles to articulate a failure-sensitive bridge between personal-level philosophical roles and sub-personal computational roles (Section 1.4). For the key constructive moves—theory as admissibility constraints on model space and observation statements as norm- and instrument-sensitive report policies—we refer to Sections 1.5, 1.7, and 4.3; here we only situate the contribution relative to the Bayesian PoS toolbox.

Much Bayesian philosophy of science proceeds by taking the hypothesis language (or model class) and the measurement/likelihood interpretation as fixed, and then asking how evidence confirms hypotheses, how priors should be chosen or constrained, and how model comparison trades off fit and complexity. On that standard picture, theory-ladenness is often discussed in terms of prior dependence or model-choice dependence within an already articulated space. Our claim is compatible with those tools, but it shifts the explanatory locus: the central Hanson-style question is not merely how to *weight* hypotheses, but how scientific theorizing helps *constitute and stabilize* what counts as an eligible variable, an admissible structure, and a licensed measurement family in the first place.

This shift clarifies why "theory = prior" is (at best) a potentially misleading slogan. Even when admissibility restrictions can be represented as priors *formally*, their methodological role is different: priors regulate updating *within* a fixed inferential setting, whereas theories (in our narrow sense) function as admissibility constraints that shape the space in which priors and likelihoods are well-defined (Sections 1.5 and 1.7). Moreover, because scientific observation is a public, norm-governed practice, we model observation statements as report policies constrained by communal norms and instrumentation/calibration regimes (N, I) as well as local goals and contexts (G, C) (Section 4.3). Finally, predictive processing is used here not as a normative foundation but as a sub-personal role vocabulary that makes the bridge explicit and failure-sensitive (Section 1.4): the point is to articulate level-coherent role correspondences, not to identify public theories with neural priors.

The novelty of the present proposal is threefold. First, it treats theories as *meta-level admissibility constraints* on model space (Sections 1.5 and 1.7), thereby locating "theory-ladenness" primarily in how hypothesis spaces, structures, and measurement families are licensed [12, 29, 40]. Second, it models public observation statements as *report policies* over posterior states under explicit communal constraints (N, I) and local pragmatics (G, C) (Section 4.3) [2, 11]. Third, it recasts discovery dynamics as multi-tier revision with explicit success criteria and failure modes (Section 1.4), and connects anomaly diagnosis to standard model-checking practice via structured residuals (Section 1.12) [3, 10, 12, 26].

Methodologically, our aim is not to offer a purely normative Bayesian reconstruction of scientific inference, but to articulate a *bridge hypothesis* between personal-level philosophical roles (theory, observation, anomaly, discovery) and sub-personal computational roles (generative inference, precision-weighted

error, model updating/selection). The bridge is therefore explicitly *failure-sensitive*: if the proposed role-constraints do not hold, the correspondence fails (Section 1.4).

1.11 What the Bridge Adds: Explanatory Payoffs Beyond Restatement

A natural worry is that a PP/Bayesian reconstruction merely *restates* Hanson in new vocabulary. The payoffs below are intended to be substantive and operational, given the bridge criteria and failure modes in Section 1.4.

1. **A principled locus-of-variation story (precision).** The bridge predicts *where* theory-ladenness is likely to be "hard" vs. "soft": when low-level sensory signals are assigned high precision, top-down influence is limited; when precision is redistributed to higher levels, seeing-as becomes more theory-sensitive [3, 10, 20]. This yields a non-binary alternative to simple penetration/non-penetration dichotomies [9].

2. **Two-tier dynamics with an explicit failure criterion.** The bridge makes explicit when anomalies should be treated as resolvable by within-model tuning and when they should trigger between-model selection: persistent *structured* residuals under best within-model fit (especially under high precision) function as a computational proxy for "anomaly" [12, 26].

3. **A minimal model of observation reports as actions.** By treating observation statements as report policies over posterior states, the framework explains why identical sensory input can yield systematically different observation reports under different goals, contexts, and community norms, without collapsing the issue into a mere "probability-to-text" mapping [2, 11].

4. **A clarified target for "theory" (meta-constraints).** The bridge makes explicit *where* "theory" enters the Bayesian/PP picture: not as a single numerical prior, but as constraints on admissible model space (variables, structure, likelihood/noise licenses, and durable precision expectations). For the minimal specification and the "constraints-not-priors" argument, see Sections 1.5 and 1.7 [26, 29, 40].

These payoffs are modest but substantive: they sharpen Hanson's insights into a set of testable and failure-sensitive structural claims.

1.12 Operational Discriminators and Empirical Touchpoints

To keep the bridge hypothesis from being a mere relabeling, we state three operational discriminators (D1–D3) that connect the proposed role mappings to empirical and methodological patterns.

D1 (Precision manipulation) *Prediction:* the degree and locus of theory-ladenness should vary with precision allocation [3, 10, 20]. When low-level sensory channels are assigned high precision, top-down theoretical influence should be limited (the "hard" case); when precision is redistributed toward higher-level hypotheses (e.g., under ambiguity, degraded input, or task-driven uncertainty), seeing-as and categorization should become more theory-sensitive (the "soft" case). This discriminator provides a graded alternative to binary penetration claims [9].

D2 (Structured residuals vs. noise) *Prediction:* anomalies that rationally motivate *between-model* shifts should correspond to *persistent, structured* residual patterns that survive best within-model tuning under an appropriate noise/measurement model [12, 26]. In contrast, discrepancies that dissipate under plausible changes to precision weighting, attention, or measurement/noise assumptions are expected to remain within-model (including "instrument-model" repair) rather than triggering model-class revision [2].

Practical diagnostics (toy checklist). In ordinary statistical practice, "structured residuals" are diagnosed by a small family of robustness and model-checking routines rather than by a single magic criterion. For present purposes, the following checklist is sufficient:

- **Posterior predictive checks:** compare replicated data from $p(\tilde{s} \mid M)$ (or $p(\tilde{s} \mid s, M)$) with observed s; look for systematic discrepancies rather than isolated outliers [12].
- **Residual structure:** test residuals for autocorrelation, periodicity, and systematic bias (e.g., patterning by time, condition, or instrument setting).
- **Noise-model robustness:** check whether the pattern survives plausible likelihood variants (e.g., heavy-tailed errors, heteroscedasticity, correlated noise, or mixture/outlier components) [2, 12].
- **Calibration/processing sensitivity:** verify that the residual pattern is stable across reasonable calibration or preprocessing revisions (linking to I rather than to world-model structure).
- **Cross-setup stability:** where applicable, test whether the pattern reproduces across instruments/labs under comparable reporting and calibration regimes (N, I) [2].

Controlled circularity (not a vicious one). What counts as a "plausible" family of noise/measurement variants is indeed theory- and practice-dependent (via the licensing package $\mathcal{L}_T$ and its instance-level instantiations in I). The proposal therefore does not pretend to eliminate circularity; it *manages* it. The key external constraints are communal: cross-instrument replication, calibration standards, and inter-lab robustness (I, N) restrict which re-specifications count as acceptable repairs, and they stabilize when residual structure is treated as methodologically significant rather than dismissible noise [2, 12]. When persistent residual structure survives these managed checks, pressure can shift to the licensing package itself (including revision of $\mathcal{L}_T$), which is exactly the failure-sensitive loop the bridge is designed to capture.

D3 (Report-policy dissociation) *Prediction:* systematic variation in public observation statements can be induced *without* altering perceptual content by manipulating reporting goals, thresholds, and community constraints (parameters G, C, N, I), holding stimuli and core inferential constraints fixed. Conversely, manipulations that genuinely alter priors/precision in the underlying generative model should change posterior structure and thus propagate to both seeing-as and report behavior [9, 11]. This discriminator operationalizes the distinction between perceptual change and post-perceptual/report-level change without presupposing a strong cognitive-penetration thesis.

Mini-case (how N shapes reports and how I stabilizes them). Consider a detection-style setting in which the same calibrated data stream s is analyzed under a fixed generative model and yields a stable posterior over a latent "signal-present" variable z. Suppose the posterior support for $z =$ signal is high but not decisive for a categorical claim. Even *holding the posterior structure fixed*, different

community norms N can rationally force different public observation statements: one community may require an exceptionally stringent evidential threshold for the speech act "detect" (and mandate explicit uncertainty reporting), thereby licensing only a hedged report (evidence) or an upper-bound; another community, with different historical error-cost asymmetries and reporting conventions, may treat the same posterior as sufficient for a categorical detection claim. Likewise, the same posterior can support different report choices when the goal/context (G, C) changes (e.g., archival reporting vs. real-time warning), even when the perceptual/inferential state is unchanged.

Crucially, this is not a subjectivist loophole. The norms N are publicly enforceable constraints on what counts as an acceptable observational commitment, and instrumentation/calibration regimes I constrain which posterior states are even reproducible across laboratories. In mature practice, what stabilizes as a "fact" is therefore not a private posterior as such, but a *report type* that is robust under shared I and N (and under reasonable model-checking and noise-model variants), mirroring the data/phenomena distinction emphasized in the philosophy of measurement and modeling [2, 12]. This is the intended content of D3: systematic report variation can be induced by policy constraints (G, C, N, I), without requiring a change in perceptual content, while still remaining methodologically constrained and publicly assessable.

These discriminators sharpen the failure modes stated in Section 1.4: D3 without D1/D2 is a natural candidate for **F1**, whereas D2 without a corresponding reorganization of seeing-as is a natural candidate for **F2**.

Accordingly, Sections 2 and 3 are intentionally selective: they are not comprehensive surveys, but targeted reconstructions that serve the bridge claims developed in Sections 4–5.

1.13 Two Worked Micro-Cases: How D1–D3 Show Up in Scientific Practice

To illustrate that D1–D3 are not merely labels, we flag two recurring patterns (sketched only at a high level). First, in detection-style measurement contexts, high effective precision (high SNR and stable calibration) tends to limit top-down influence on posterior structure (D1), while low effective precision increases prior- and model-sensitivity; nevertheless, public reports can still vary with reporting goals, thresholds, and community norms even when posterior structure is held fixed (D3). Second, strong "anomaly" diagnoses in mature practice typically track the persistence of *structured* residual patterns across best within-model tuning and plausible measurement/noise-model repairs (D2). We return to these patterns in Sections 4–5, where the multi-tier dynamics are developed in detail.

1.14 Structure of the Paper

The remainder of the paper is organized as follows.

1. Section 2 reconstructs Hanson's account of observation and extracts the relations among observation, data, theory, and discovery at the personal level.

2. Section 3 introduces the predictive-processing framework and the Bayesian brain hypothesis, outlining the relevant computational structure at the sub-personal level.

3. Section 4 presents a computational reconstruction of the theory-ladenness of observation and develops an integrated model based on the idea of theory as a meta-structure that constrains high-level priors and the admissible model space.

4. Section 5 analyzes Kepler's discovery as a case study, distinguishing between incremental updating characteristic of normal science and model-selection processes characteristic of paradigm shifts [38].

5. Finally, Section 6 summarizes the overall argument and discusses the implications and limitations of bridging the personal and sub-personal levels.

2 Theory-Ladenness of Observation in Hanson

This section offers an analytical reconstruction of N. R. Hanson's account of the theory-ladenness of observation as developed in *Patterns of Discovery* (1958) [16]. Hanson does more than merely assert that observation is "not independent of theory." He articulates a network of interrelated claims that together structure what he calls observation in mature science: (1) the relation between *seeing* and *seeing-as*, (2) the linguistic dependence of observational reports, (3) the conditional and practice-bound character of "facts", (4) the theory-ladenness of causal explanation, and (5) the structure of hypothesis formation and discovery.

Deliverable of this section. Our goal here is to extract a *personal-level role map* that can later be put in explicit (but non-reductive) correspondence with a predictive-processing and Bayesian vocabulary. Accordingly, the output of this section is a structured characterization of (i) seeing-as, (ii) observation statements as public, truth-evaluable reports, (iii) fact-stabilization as a practice-bound achievement, (iv) causal explanation as framework-relative selection of dependencies, and (v) discovery as reorganization driven by anomaly.

Hanson's project forms a central component of the mid-century reorientation of philosophy of science toward historically and psychologically realistic accounts of scientific practice, alongside Kuhn's paradigm theory and Feyerabend's methodological critique [8, 24]. It also resonates with Sellars' critique of the "Myth of the Given," which denies that perceptual episodes can function as justificatory givens apart from their conceptual roles [32]. Finally, Hanson's concerns anticipate later probabilistic and model-based approaches that re-examine the relations among data, evidential support, and theoretical structure [12, 21, 26].

Importantly, the aspects of Hanson's analysis examined in this section belong to what Section 1 called the *personal level*—the level of scientists' reasoning, judgment, and linguistic practice. By contrast, the predictive-processing framework introduced in Section 3 targets the *sub-personal level* of neural and computational mechanisms [5, 27].

The objective of this paper is not to collapse these levels into a reductive identity. Rather, drawing on Marr's levels of analysis [27], Dennett's relations among the intentional, design, and physical stances [7], and mechanistic accounts of multi-level explanation [5, 22], the aim is to clarify *structural correspondences* between explanatory roles at the personal level and computational roles at the sub-personal level. In line with the criteria stated in Section 1.4, the correspondence is not a strict mathematical isomorphism but a role- and constraint-preserving mapping: it should illuminate what drives revision (anomaly), what forms of revision are available (within-model adjustment vs. model-class change), and what is reorganized when revision is radical (the space of admissible observations, facts, and explanations).

Accordingly, the purpose of this section is not to catalogue Hanson's remarks as a series of isolated episodes, but to extract from his discussions a personal-level model of the relations among *observation*, *data*, and *theory*. With an eye toward the PP connection developed in Section 3, the following subsections examine each component of this structure in turn.

2.1 The Structure of *Seeing* and *Seeing-as*

At the core of Hanson's analysis lies the claim that what matters epistemically in scientific observation is not merely the reception of sensory stimulation, but an organized perceptual uptake of the world *as* containing objects, events, and processes—that is, *seeing-as* [16]. What is often treated as a simple "raw" visual experience is, for Hanson, something whose structure must itself be interrogated.

To motivate this point, Hanson introduces the contrast between Kepler and Tycho Brahe. Confronted with the same visible scene (e.g., the apparent diurnal motion at sunrise), they agree on the appearance but organize and describe it under different kinematic commitments: for Kepler, the observed pattern is naturally taken as compatible with a moving Earth; for Tycho, the Earth remains at rest and the same appearance is organized under a different cosmological framework [16, 38]. Hanson's point is not a claim about differential retinal stimulation; it is a claim about what the observers are prepared to *take* the visible scene to be an instance of.

For expository purposes, it is helpful to distinguish two layers that Hanson treats as importantly different in *scientific* observation:

- **Sensory registration ("seeing" in a thin sense)**: stimulus-driven uptake and early processing that (at least in many ordinary cases) can function as comparatively shared constraints across observers and that constrains what can be seen at all. (This is not claimed to be wholly theory-free; rather, it is not where Hanson locates the explanatory differences of primary interest.)
- **Interpretive uptake ("seeing-as")**: organization of the stimulus as an object, event, or process, under concepts and background commitments that determine what is salient and what counts as relevant structure [16].

This distinction becomes vivid in classical Gestalt ambiguity, such as the "old woman/young woman" figure (Fig. 1) [19]. Here the proximal stimulus is held fixed, yet perceivers can undergo a reorganization in which the same lines are taken *as* an elderly woman or *as* a young woman. What changes is not the stimulus but the organization of the stimulus into a meaningful whole. Hanson's interest is not the psychological curiosity of illusions; it is the structural moral that mature scientific observation likewise involves organization under learned conceptual and theoretical schemes [6, 16].

Hanson reinforces the same point with mundane examples: a written leaflet may appear to a non-speaker as a mere array of marks, while to a fluent reader it is immediately seen *as* meaningful text. Again, the stimulus can be shared while the uptake differs.

It is useful to note a family resemblance between Hanson's "seeing-as" and later cognitive-scientific slogans such as Gregory's "perception is a hypothesis" [15], as well as Bayesian and predictive approaches to vision in which perceptual content is understood as inference to latent causes under a generative model [23, 25]. Hanson does not offer such a computational account; his contribution is to isolate, at the personal level, the role that organization and background commitment play in observation.

Figure 1: The ambiguous "old woman / young woman" figure. Based on Hill (1915) [19].

As Section 3 will show, predictive-processing models provide a way to revisit this topic by distinguishing sub-personal processes close to sensory input (corresponding to the thin sense of seeing) from higher-level generative-model-based processes (corresponding to seeing-as), with their relative influence modulated by precision assignment [3, 10, 20]. The present paper emphasizes a cross-level structural correspondence between these descriptions without reducing Hanson's personal-level analysis to a sub-personal mechanism.

In sum, Hanson's basic insight is that observation is not merely the registration of retinal images but the interpretive uptake of the world through conceptual and theoretical frameworks [16].

2.2 Linguistic, Conceptual, and Contextual Dependence: The Non-neutrality of Observation Statements

While the distinction between *seeing* and *seeing-as* already shows that observation involves organization, Hanson further emphasizes that scientific observation is inseparable from language [16]. Scientists typically communicate observational outcomes in the form of *observation statements*. For Hanson, it is precisely at this linguistic level that theory-ladenness becomes especially salient.

A central contrast is that observation statements are truth-evaluable: they have truth conditions and can function as premises in inference, whereas a picture or sensory presentation is not truth-evaluable *in the same way* as a proposition is [16]. To transform an observational episode into a reportable claim, the world must be articulated into subjects and predicates, and such articulation involves conceptual and syntactic choices.

Consider the image in Fig. 2.

We normally describe it as "The bear is climbing the tree." Although it is logically possible to say "The tree is lifting the bear upward," our background knowledge about biology and mechanics renders the latter description pragmatically and explanatorily inappropriate. The point is not that language merely decorates a neutral visual core, but that reporting an observation as a candidate fact already presupposes a shared conceptual organization.

Context-sensitivity in ordinary language underscores the same lesson. The utterance "Fire!" can mean "Shoot!" on a training ground but "There is a fire!" in a theater. Such examples serve Hanson to show

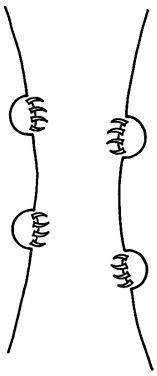

Figure 2: A bear climbing a tree. Redrawn from Hanson (1958) [16].

that the meanings of observation words and observation statements depend on contexts of action and communication, and thus cannot be treated as a purely neutral "observation language" [16].

Here Hanson's position aligns with Sellars' critique of the Myth of the Given: to play epistemic roles, perceptual episodes must be situated within conceptual and linguistic practices [32]. In later sections, this linguistic point will be connected to the bridge framework developed in Section 1: public observation reports will be modeled as *report policies* that extract actionable public commitments from an agent's epistemic state under goals and contexts, rather than as direct readouts from raw data [2, 12].

2.3 Facts and Data: Stabilization, Measurement, and Practice-Bound "Facts"

In everyday use, the term "fact" suggests something stable, universal, and independent of any particular observer. Hanson challenges this naïve conception by emphasizing that what counts as a "fact" in scientific practice is often conditional on background frameworks of description and on the context of use [16]. The point is not relativism about truth, but that the *conditions of fact-stating*—what is taken to be settled enough to assert—depend on conceptual resources, measurement practices, and shared standards.

A central locus of such dependence is *measurement and stabilization*. Many scientific "facts" are not raw data points but stabilized report types secured through calibration, error modeling, and communal reporting norms. For example, whether a time series licenses a public claim of *detection*, a hedged claim of *evidence*, an *upper limit*, or a *non-detection* depends not only on the recorded signal but also on background assumptions about noise, instrument behavior, and acceptable thresholds for assertion (cf. the data/phenomena distinction) [2]. Similarly, whether a discrepancy is treated as a measurement artifact, as noise, or as a robust phenomenon can depend on how error bars are modeled, how calibration procedures are justified, and how results are checked across laboratories.

A related issue concerns categorization. Predicates such as "red" or "blue" presuppose a classificatory scheme and standards for applying terms; likewise, scientific categories (e.g., "signal," "artifact," "background," "outlier") encode learned conceptual resources and practical criteria for application. The point is not that such predicates lack any physical basis, but that the space of expressible and reportable "facts" depends on conceptual and methodological capacities [16, 32].

To connect Hanson's point to the bridge framework of Section 1, it is helpful to distinguish *data*

from *phenomena*. Although Hanson does not employ this later terminology, Bogen and Woodward's distinction clarifies how theory-ladenness can coexist with objectivity: data are often local, instrument- and practice-dependent records, whereas phenomena are comparatively stable patterns extracted from those data across methods and contexts [2]. On this reading, "facts" in mature science are rarely bare data; they are more often stabilized, communicable report classes supported by shared constraints on measurement, interpretation, and reporting.

This point will matter later when we introduce a computational vocabulary: the paper will treat "facts" not as theory-free atoms but as stabilized outcomes of interpretive and methodological practices under shared admissibility constraints and measurement regimes.

2.4 Causal Chains: The Theory-Ladenness of Explanation and Inference

Hanson also argues that causal explanation depends on the observer's theoretical framework [16]. A stream of events does not wear its causal structure on its sleeve; rather, scientific explanation involves selecting variables, carving event-streams into relevant units, and organizing them into explanatory relations. Accordingly, different theoretical perspectives can pick out different causal "chains" or explanatory linkages from the same underlying sequence.

Hanson's moral can be expressed in a way that connects naturally to later formal work on causality: causal claims are evaluated within representational frameworks that specify which variables are candidates for causes, which structural relations are admissible, and which interventions would make a difference [29, 40]. Read in this light, Hanson anticipates (at the personal level) the idea that causal structure is not simply read off from uninterpreted data, but depends on modeling assumptions that determine what counts as a cause, an effect, and an explanatory pathway. Later sections will use this connection to motivate the treatment of theory as a meta-level constraint on admissible model space, rather than as a mere prior distribution (Section 1.5).

A clarifying distinction (preview). In what follows, "causal structure" is primarily invoked in this *framework-relative, explanatory-role* sense. A further question—central to interventionist accounts—is when such dependency structures license claims about interventions and counterfactual difference-making. We return to this distinction later when we introduce the generative-model sense of "causality" used in predictive-processing formulations.

2.5 Hypothesis Formation and Pattern Recognition: The Structural Features of Discovery

For Hanson, the theory-ladenness of observation is not a skeptical thesis about distortion. It is a condition that makes scientific discovery possible. Discovery is not the mechanical accumulation of data but an active process in which observers detect and stabilize patterns by deploying and revising hypotheses [16].

Kepler's reformulation of planetary motion provides a canonical illustration. Working within prevailing circular-orbit commitments, Kepler organized and reorganized observational results while confronting persistent misfits that could not be eliminated by routine adjustments. Through a sequence of conjectures and refinements—including explorations of non-circular trajectories and revisions of auxiliary assumptions—Kepler arrived at the elliptical model as a superior way of organizing the relevant data [6, 38]. The point is not that data uniquely determine a hypothesis, but that hypotheses restructure what counts as salient

structure in the data, thereby enabling new phenomena to become visible.

Hanson connects the logic of such reorganizations to Peirce's notion of abduction (hypothetical inference) [30]. Abduction is not a mysterious leap from data but a sensitivity to latent patterns that become salient under an appropriate conceptual organization. This perspective aligns with later defenses of inference to the best explanation [17].

These structural features also foreshadow the role of generative models in the predictive-processing frameworks introduced in Section 3. In PP, hypotheses correspond (at a sub-personal level) to internal generative models; sensory inputs are received and interpreted *as* instances under those models, and persistent precision-weighted prediction errors can drive updating or model-class revision [10, 11, 26, 31]. Later sections will exploit this parallel to map Hanson's observation–anomaly–discovery pattern onto Bayesian updating and model comparison dynamics.

2.6 Summary: Hanson's Personal-Level Role Structure and Its Bridge-Relevance

The foregoing analysis allows Hanson's account of observation to be summarized in three theses:

1. Observation is not mere sensory registration but *seeing-as*, and is therefore structured by the observer's concepts, background commitments, and training [16].

2. Observation statements, and the "facts" stabilized through scientific practice, are linguistically and contextually articulated; the ideal of a theory-neutral observation language is difficult to sustain [16, 32].

3. Discovery is an active process in which observers reorganize experience through hypotheses, including abductive and explanatory forms of inference [16, 17, 30].

Hanson thus presents theory-ladenness not as a purely skeptical thesis but as a positive condition of possibility for scientific practice [16]. At the same time, his analysis remains primarily descriptive and conceptual. It does not specify sub-personal mechanisms that realize seeing-as, nor does it provide a computational criterion for when anomalies rationally warrant model-class change.

This paper therefore seeks to map the personal-level structure identified by Hanson onto the sub-personal framework of predictive processing. In line with the bridge criteria of Section 1.4, the correspondence of interest can be summarized as follows: (1) tension between input and expectation functions as a driver; (2) surprise/anomaly triggers revision; and (3) revision includes both within-model adjustment and, in some cases, between-model selection that reorganizes the space of admissible observation and explanation [11, 24, 26].

Finally, regarding how theory intervenes in observation, this study rejects the simplifying slogan *theory = prior*. Instead, theory is understood as a meta-level structure that constrains which variables may count as candidate causes, and which causal and likelihood structures are admissible within the model space (cf. Section 1.5) [2, 29, 40]. Section 3 develops predictive processing and the Bayesian brain hypothesis as the sub-personal framework for addressing the explanatory gap left open by Hanson's personal-level account.

3 Predictive Processing and the Bayesian Brain Hypothesis

In Section 2, we reconstructed Hanson's account of theory-ladenness at the *personal level*: scientific observation is not a neutral reception of sensory input but a concept- and context-dependent process of *seeing-as*. Yet Hanson's account remains primarily conceptual and historical; it does not specify the sub-personal cognitive mechanisms or computational structure that could realize such theory-ladenness.

This section begins to fill that gap by surveying the contemporary cognitive-neuroscientific framework of *predictive processing* (PP) and the *Bayesian brain hypothesis*. Very roughly, PP treats perception and learning as forms of *approximate inference* under hierarchical generative models, in which mismatch signals (prediction-error-like discrepancies or closely related variational quantities) play a functional role in belief updating and, in some formulations, action selection [3, 10, 11, 20, 31].

Deliverable of this section. Our aim here is not to defend one neuroscientific doctrine, but to extract a *sub-personal role map* that can later be put into explicit correspondence with the personal-level roles reconstructed from Hanson. The output of this section is therefore a structured characterization of (i) hierarchical *generative inference* (posterior estimation under a model), (ii) *precision-weighted mismatch* as a revision-driving signal, and (iii) *model evidence and model comparison* as resources for between-model selection.

Methodological stance (minimal commitments). The bridge developed here does *not* presuppose that the free-energy principle (FEP) or a strong "Bayesian brain" thesis is established as a complete neurobiological theory. What we require is weaker: (i) perception and learning can be fruitfully modeled *as if* they involved approximate inference under hierarchical generative models, (ii) mismatch (prediction-error-like) signals play a functional revision-driving role, and (iii) uncertainty/precision modulates the influence of top-down expectations. If even these weak commitments fail, the proposed correspondence fails—which is why we formulated explicit success conditions and failure modes in Section 1.4 (see also Subsection 3.8 below).

A crucial point must be emphasized at the outset: the PP models discussed here operate at the *sub-personal level*, describing information-processing in the brain. They are not to be conflated with the *personal-level* descriptions of scientists' reasoning and linguistic practice analyzed in Section 2 [5, 27]. The aim of the present work is to identify a computational and structural correspondence (a *realization relation*) between these levels, in the modest sense articulated by the bridge criteria in Section 1.4. Thus, the central question for the remainder of this section is:

> *What role- and constraint-preserving correspondences can be identified between the personal-level relations among observation, anomaly, and discovery described by Hanson and the sub-personal relations among generative inference, (precision-weighted) prediction error, and model updating/selection posited by predictive processing?*

Terminological note. In what follows, *predictive coding* refers to a family of mechanistic architectures often discussed in relation to hierarchical cortical signaling; *predictive processing* refers to the broader research program; and *active inference* and the *free-energy principle* (FEP) refer to a variational formalism

and its extensions that unify perception and action within a single inferential framework [4, 11]. These labels are sometimes used interchangeably; here we keep them distinct for clarity.

Outline. The section proceeds as follows.

1. Section 3.1 introduces the intuitive picture of predictive coding.
2. Section 3.2 reviews the Bayesian inferential framework that underwrites these models.
3. Section 3.3 examines generative models in relation to hierarchical cortical organization.
4. Section 3.4 introduces prediction-error minimization and the free-energy principle.
5. Sections 3.5–3.6 discuss PP interpretations of perceptual construction, including Seth's "controlled hallucination" slogan and the role of precision in illusions and psychopathology.
6. Section 3.7 synthesizes these elements to articulate a computational account of what it means for the brain to "see" the world.
7. Finally, Section 5 will use these tools to reconstruct Hanson's account of discovery as a two-tier process of within-model updating and between-model selection.

3.1 Predictive Coding as a Model of Perception

Predictive processing reconceives perception not as passive registration of sensory inputs but as an ongoing cycle of prediction and correction. Early hierarchical models of predictive coding are exemplified by Rao and Ballard's account of visual cortex [31], later generalized by Friston into a theory of cortical responses and the free-energy principle [10, 11].

The basic ideas may be summarized as follows:

- The system maintains an internal model of hidden causes of sensory signals (objects, events, latent variables), and uses this model to generate predictions about forthcoming sensory input.
- Incoming sensory signals are compared with these predictions, and the discrepancy is encoded (often implicitly) as a *prediction error*.
- The system reduces such discrepancies by updating internal states (beliefs) and, in some formulations, by acting so that sampled sensory input better conforms to predictions (active inference).

Crucially, sensory inputs are not simply fed upward through a hierarchy. Rather, higher-level areas send top-down predictions to lower levels, while lower levels send bottom-up mismatch signals reflecting deviations from those predictions. This bidirectional architecture—top-down *predictions* and bottom-up *errors*—is the signature of predictive coding [10, 31].

Where traditional feedforward accounts conceive perception as a largely unidirectional cascade from receptors to higher cognition, predictive coding instead emphasizes a continual cycle of hypothesis generation and error-driven correction. Perception becomes, in a computational sense, an ongoing form of hypothesis testing [3, 20].

It is important not to overstate empirical conclusions here. The present goal is not to review the experimental evidence for any specific implementation, but to extract the *role structure* that will later be put in level-coherent correspondence with Hanson's personal-level account: top-down expectations, bottom-up mismatch signals, and revision dynamics modulated by uncertainty/precision [10, 11].

3.2 Bayesian Inference: Prior, Likelihood, Posterior, and Model Evidence

The mathematical backbone of many PP formulations is *Bayesian inference* and its approximate implementations. This subsection reviews the core components of Bayesian updating—priors, likelihoods, and posteriors—as well as the notion of *model evidence* that will later matter for between-model selection [12, 21, 26].

For simplicity, suppose we have observed a datum d and wish to update beliefs concerning a discrete set of candidate hypotheses h. Prior beliefs about h are encoded in the *prior* $p(h)$. The probability of observing d given h is the *likelihood* $p(d \mid h)$. The updated belief after observing d is the *posterior* $p(h \mid d)$.

Bayes' rule expresses their relationship:

$$p(h \mid d) \propto p(d \mid h)\, p(h),$$

that is,

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h'} p(d \mid h')\, p(h')}.$$

In this framework, priors encode background expectations; likelihoods encode a hypothesis-to-data mapping; and posteriors encode the updated inferential state in light of the data. In many cognitive and neural applications, such updating is understood as approximate and sequential rather than exact and batch-style [4, 11].

A further distinction is crucial for the present paper. Bayesian updating presupposes a hypothesis space (or model class) over which the update is carried out. When competing model classes M are in play, Bayesian model comparison appeals to *model evidence* (marginal likelihood) $p(d \mid M)$ and the posterior over models:

$$p(M \mid d) \propto p(d \mid M)\, p(M),$$

where

$$p(d \mid M) = \sum_{h} p(d \mid h, M)\, p(h \mid M)$$

(or the corresponding integral in continuous cases) [1, 12, 26]. This distinction between within-model updating and between-model selection will be central later for reconstructing discovery as a two-tier process.

Using Hanson's terminology from Section 2, what we call "theory" or "paradigm" maps not to a single numerical prior but to a higher-level structure that constrains which priors, likelihood forms, and model architectures are admissible. Accordingly, rather than reducing a theory to a particular prior distribution, this paper interprets theories as *meta-level constraints on model space*: constraints on candidate variables, admissible causal structures, and likelihood/measurement assumptions (Section 1.5) [2, 29, 40]. Thus the operative correspondence is not *theory* = *prior*, but:

theory constrains which priors and generative-model architectures are admissible.

The Bayesian framework is fundamental in statistical pattern recognition and probabilistic machine learning [1, 28]. Predictive processing can often be fruitfully modeled *as if* it implemented approximate Bayesian inference: the system maintains expectations about latent causes, incorporates sensory inputs via a likelihood (measurement) model, and updates internal states in ways that reduce mismatch signals under precision control [3, 10, 11, 20]. This is the core idea behind the Bayesian brain hypothesis (under the present, minimal, modeling-oriented reading).

3.3 Generative Models and Hierarchical Cortical Architecture

In Bayesian inference, the likelihood specifies how probable the observed data would be if a hypothesis about hidden causes were true. If a hypothesis is interpreted as an internal representation of some latent cause in the world, then the likelihood specifies how that cause generates the sensory data. The corresponding internal structure is a *generative model* [1, 26].

Formally, a generative model M can be expressed as a joint distribution

$$p(s, z \mid M) = p(s \mid z, M)\, p(z \mid M),$$

where z denotes hidden causes (latent states) and s denotes sensory signals. Here $p(z \mid M)$ is a prior over hidden causes, and $p(s \mid z, M)$ is a measurement/likelihood model describing how causes generate sensory input.

Predictive-processing models often propose that this generative structure is mirrored (at least approximately) in hierarchical cortical organization. Lower sensory cortices represent comparatively local, short-timescale features, while higher areas encode increasingly abstract and integrative structure. In a hierarchical generative model, higher-level states generate the states below them, culminating in sensory signals at the lowest level. This is often represented by a cascade of conditional distributions $p(z_{l-1} \mid z_l, M)$ [23, 25].

Top-down signals convey predictions—"if the world is in state z, then the sensory input should look like s"—while bottom-up signals convey mismatch signals (prediction errors) computed by comparing predicted and actual sensory input. These mismatch signals are then used to update both the hidden states and, in learning contexts, parameters of the generative model.

When juxtaposed with Hanson's analysis in Section 2, the significance of hierarchical generative structure becomes clear. Higher levels of a generative model can serve as computational counterparts to relatively abstract background expectations, while intermediate levels encode categorical and schematic organization, and lower levels remain more tightly constrained by sensory input. This is *not* a one-to-one identification of personal-level "theories" with particular neural representations; it is a structural correspondence that supports the bridge project of maintaining level consistency without reduction [5, 27].

3.4 Prediction-Error Minimization and the Free Energy Principle

Within predictive processing, mismatch signals are not treated as mere computational residuals; they are treated as quantities that drive belief updating (and, in some formulations, action selection). For expository purposes, prediction error is sometimes written schematically as "actual minus predicted." However, in

standard PP formulations, mismatch signals are typically *precision-weighted* and defined across multiple hierarchical levels. A minimal schematic expression is therefore:

$$\varepsilon \;\approx\; \Pi\,(s - \hat{s}),$$

where Π denotes (possibly level-dependent) precision (inverse variance) weighting [10, 11]. In standard accounts, such mismatch terms are distributed across hierarchical levels and can concern both hidden-state estimates and (in learning contexts) parameters of the generative model.

A more general formalization is given by Friston's *free-energy principle* (FEP) [11]. Let $q(z)$ denote an approximate posterior (an internal variational representation) over latent causes z under a model M. The variational free energy can be written as

$$F[q] \;=\; \mathrm{E}_{q(z)}[-\log p(s, z \mid M)] \;+\; \mathrm{E}_{q(z)}[\log q(z)],$$

and one can show the standard variational identity

$$F[q] \;=\; \mathrm{KL}\big(q(z) \,\|\, p(z \mid s, M)\big) \;-\; \log p(s \mid M),$$

which implies

$$F[q] \;\geq\; -\log p(s \mid M).$$

Thus minimizing $F[q]$ both (i) makes $q(z)$ a better approximation to the posterior and (ii) tightens an upper bound on sensory "surprise" (negative log evidence) under the model [11, 26].

In PP/FEP frameworks, perception corresponds to updating internal states (or beliefs) so as to reduce free energy, while action can be understood as selecting policies that change the sampled sensory input so as to reduce free energy. This extension to action is often referred to as *active inference* [4, 11].

The present paper does not require the full technical apparatus of FEP. What matters for the bridge is the role structure: hierarchical generative models, mismatch signals that drive revision, and precision/uncertainty as a parameter controlling the balance between top-down expectations and bottom-up signals. From this perspective, what Hanson called theory-ladenness of observation is exactly the kind of phenomenon expected when top-down models and precision-weighted mismatch signals jointly shape perceptual content.

Finally, note how the variational identity connects to *model evidence*. Since $-\log p(s \mid M)$ is the negative log evidence, minimizing free energy can be interpreted as a process that (approximately) increases evidence by improving posterior approximation. This is a *formal* connection, not a normative vindication of scientific method: the present paper does not claim that scientific inference is rational *because* brains minimize free energy [12, 26]. Accordingly, FEP/PP is treated as a naturalized candidate framework that can support a level-coherent bridge, while normative assessment remains a distinct philosophical task [18, 40].

3.5 Anil Seth's "Controlled Hallucination" and the Role of Precision

A vivid formulation of PP's philosophical implications is Anil Seth's claim that *perception is a controlled hallucination* [34]. Read charitably (and stripped of rhetorical excess), the slogan highlights a core PP commitment: perceptual content is generated by top-down predictions under a generative model, while

sensory signals—weighted by their estimated precision—function as error-correcting constraints that keep perception tethered to the environment [3, 10, 20].

On this view, perception is not a direct copy of external reality but an actively constructed, hypothesis-like representation continuously corrected by incoming evidence. The qualifier "controlled" matters: precision-weighted mismatch signals determine the degree to which predictions dominate perceptual content and the degree to which sensory input can override those predictions.

Juxtaposed with Section 2, Seth's slogan can be read as a sub-personal analogue of Hanson's theory-laden *seeing-as*: an observer sees the world *as* something under a generative model whose high-level expectations and precision assignments shape perceptual interpretation [4, 11, 20].

3.6 Illusions, Hallucinations, and Delusions: Priors, Likelihoods, and Precision

PP is often motivated by its promise to unify ordinary perception with cases of illusion, hallucination, and delusion, treating these as "natural experiments" that reveal how perceptual construction depends on priors, likelihoods, and precision allocation [15, 23, 35].

Classic illusions illustrate the Bayesian moral that perception reflects a balance between prior expectations and sensory likelihoods, modulated by uncertainty and reliability estimates. A crucial point for the bridge project is that this balance is not fixed. It is regulated dynamically by the estimated *precision* of prediction errors at different hierarchical levels [3, 10, 20]. This helps explain why some low-level illusions cannot simply be "unseen" even when one knows that the stimulus is misleading: in such cases, the relevant sensory error signals are treated as highly reliable. By contrast, in ambiguous stimuli or high-level categorization tasks, sensory signals can be treated as less precise, and higher-level expectations exert proportionally greater influence.

In psychopathology, PP-style accounts have been applied to hallucinations and delusions by emphasizing aberrant priors and, especially, aberrant precision weighting. Very roughly (and without adjudicating clinical details), hallucination-like experiences can be modeled as cases in which the precision of relevant sensory prediction errors is down-weighted relative to top-down expectations, allowing internally generated predictions to dominate experience, whereas delusion-like fixation can be associated with maladaptive precision assignment and disrupted updating dynamics [35]. The point for present purposes is structural: PP provides a principled vocabulary for locating *where* and *how* top-down expectations and bottom-up constraints differentially shape perceptual content.

This perspective also interfaces with the cognitive penetration debate. Firestone and Scholl argue that many alleged top-down effects on perception can be explained by attention, judgment, or task demands rather than by genuine changes in perceptual content [9]. PP does not, by itself, settle this debate; but it offers a graded alternative to binary penetration claims by locating variation in precision assignment and hierarchical level. In particular, stronger encapsulation is expected in regimes where low-level errors are treated as highly precise, while theory-ladenness is expected to be more pronounced in regimes where higher-level priors and precision allocation dominate seeing-as (cf. discriminator D1 in Section 1.12).

These considerations support a bridge-relevant moral: theory-ladenness is not a rare exception but a general feature of human perception, though its strength and locus vary systematically with precision and hierarchical organization.

3.7 Summary: A Computational Understanding of What It Means for the Brain to "See" the World

In this section, drawing on predictive processing and the Bayesian brain hypothesis, we outlined a computational framework for understanding what it means for the brain to "see" the world.

The discussion can be summarized as follows:

1. The brain can be modeled as maintaining hierarchical generative models that generate predictions about sensory input, with mismatch signals driving revision—an architecture often described as predictive coding and embedded within broader PP interpretations [3, 4, 10, 11, 20, 31].

2. Bayesian inference supplies the formal backbone: posterior estimation depends on priors and likelihoods, while between-model selection can be articulated via model evidence and model comparison [1, 12, 21, 26].

3. Prediction errors are typically precision-weighted and distributed across levels; precision allocation provides a principled locus-of-variation parameter that modulates the balance between top-down expectations and bottom-up signals [3, 10, 11]. This parameter will later support the bridge discriminators (D1–D3) in Section 1.12.

4. Rather than reducing theories to single numerical priors, the present account treats theories as meta-level constraints on admissible model space (variables, causal structures, likelihood/measurement assumptions, and durable precision expectations; Section 1.5) [2, 12, 29, 40].

5. PP-style interpretations of perceptual illusions and atypical cases illustrate how changes in priors, likelihood assumptions, and precision can yield systematic changes in perceptual content, thereby supplying a graded alternative to simple penetration/non-penetration dichotomies [9, 15, 35].

Predictive processing thus provides a candidate sub-personal framework for clarifying Hanson's claim that observation is theory-laden seeing-as, now articulated in computational terms. However, the correspondence between the philosophical and computational frameworks can remain merely intuitive unless it is made explicit as a role- and constraint-preserving mapping in the sense of Section 1.4. Accordingly, the next section (Section 4) advances this project by reconstructing Hanson's conceptual apparatus in the formal vocabulary of PP. There, theory-ladenness is recast as an explicit computational structure—including (i) theory as a meta-constraint on model space, (ii) seeing-as as posterior-guided inference under precision control, and (iii) observation reports as goal- and context-sensitive *report policies* rather than direct probability-to-text mappings.

3.8 Methodological Caution: PP as a Family and Alternative Realizations

Predictive processing and the free-energy principle are best read as a family of modeling strategies rather than a single, fully established neurobiological theory. For the bridge developed here, we rely only on weak commitments: (i) perception and learning can be modeled as approximate inference under generative models, (ii) mismatch signals play a functional revision-driving role, and (iii) uncertainty/precision modulates the influence of top-down expectations [4, 11].

This matters because the bridge is not a bet on one neuroscientific doctrine. Even if future work replaces current PP implementations with a different sub-personal architecture that still realizes these roles (e.g., alternative hierarchical inference schemes), the bridge claims can be reformulated at the level of roles and constraints. Conversely, if no credible sub-personal realization can be found for these roles, the bridge fails—as captured by the explicit success conditions and failure modes in Section 1.4 [5, 22]. For the corresponding clarification about public scientific theories vs. internal inferential resources (and why the bridge does not collapse levels), see Section 1.4.

4 A Computational Reconstruction of the Theory-Ladenness of Observation

Sections 2 and 3 examined, respectively, Hanson's personal-level account of theory-ladenness and the sub-personal computational framework of predictive processing. Although these bodies of work originate in distinct disciplinary contexts, they invite a shared structural reading: perception (and, in science, observation) is not mere registration but an organized uptake that is sensitive to expectations, explanatory structure, and revision in the face of mismatch [3, 10, 11, 15, 16, 20].

The aim of this section is to integrate these approaches by making Hanson's insights computationally explicit in terms of *generative models, (precision-weighted) mismatch signals, priors, posteriors, and within-model updating vs. between-model selection*. The goal is not merely to "translate" Hanson into scientific vocabulary, but to articulate a bridge hypothesis that is role- and constraint-preserving in the sense specified by the bridge criteria in Section 1.4.

Importantly, this reconstruction does *not* reduce Hanson's personal-level descriptions—concerning scientists' reasoning, discourse, and conceptual organization—to the sub-personal computations attributed to the brain in PP. Instead, the goal is to identify a *structural correspondence* (a realization relation) between these levels. In the spirit of Marr's levels of analysis [27] and Dennett's relations among stances [7], the aim is to establish a non-reductive coherence between philosophical and computational descriptions, consistent with mechanistic accounts of multi-level explanation [5, 22].

In short, this section develops a "Bayesian/PP reading" of Hanson: it maps his philosophical roles onto computational roles without collapsing the levels.

Notation and scope (roadmap). We use s for sensory/instrumental input, z for latent causes, and M for a generative-model class; posterior inference is written $p(z \mid s, M)$. A theory T is modeled (as an idealized representation of personal-level theoretical constraints) as selecting an admissible region of model space $\mathcal{M}_T \subseteq \mathcal{M}$. Public observation statements are idealized as report options u chosen by a report policy under goals and contexts (G, C) and community constraints (N for reporting norms; I for instrumentation/calibration). This notation is a modeling scaffold for the bridge, not a claim of personal/sub-personal identity.

4.1 Reinterpreting Theory as Constraints on Admissible Model Space

For Hanson, a "theory" is a high-level conceptual framework that structures how an observer sees the world: it guides *seeing-as* and shapes the background network within which observations are interpreted [16].

Similarly, Kuhn's notion of a paradigm determines what counts as a salient phenomenon and which problems and solutions are intelligible [24].

Within PP, the closest computational analogue to this personal-level role is not a single probability distribution but a family of constraints on the *generative models* under which inference is carried out. As reviewed in Section 3, generative models are often hierarchically organized, with upper levels encoding abstract hypotheses about global structure, causal organization, and invariances [11, 25].

In this paper, we follow the minimal specification from Section 1.5: a theory T is modeled as a meta-level constraint operator that selects an admissible subset of model classes,

$$\mathcal{M}_T \subseteq \mathcal{M}.$$

Importantly, this is not an exegetical claim that scientists explicitly represent T in this form. Rather, T is an *as-if* computational representation of the kinds of admissibility constraints that play the personal-level role of "theory" in Hanson's sense.

This captures the idea that theories constrain not only numerical priors but also representational resources, structural assumptions, and measurement models that make priors and likelihoods well-defined in the first place [2, 12, 29, 40].

Concretely, a theory constrains (at least) the following components:

1. which variables count as plausible *candidate causes* ($\mathcal{V}_T$);
2. which *structures* or dependency architectures are admissible ($\mathcal{S}_T$);
3. which likelihood/noise/measurement families are appropriate ($\mathcal{L}_T$);
4. which channels/levels are treated as reliable via durable precision expectations (Π_T) [3, 10].

Accordingly, the relevant dependency is not *theory = prior* but:

$$\text{Theory } T \;\rightsquigarrow\; \text{constraints on } \mathcal{M}_T \text{ (hence on admissible priors, likelihoods, and precision regimes).}$$

This avoids a misleading identification of theories with single numeric priors while still explaining how theory can shape what is taken to be observed.

This formulation also preserves the two-way constraint that motivates empirical testing rather than circularity. Theory constrains which models are admissible, but persistent, precision-weighted misfit can rationally motivate revising those admissibility constraints (cf. C1–C3 and D2 in Sections 1.4 and 1.12). In this sense, the "direction of dependence" is not one-way: theory conditions inference, and structured anomaly can force constraint change.

Finally, the Kepler–Tycho contrast can be reconstructed schematically as a difference in theory-conditioned admissible model spaces: given the same input s, different constraints yield different posteriors over latent causes,

$$p(z \mid s, M) \quad \text{with } M \in \mathcal{M}_{T_{\text{Kepler}}} \quad \text{vs.} \quad p(z \mid s, M) \quad \text{with } M \in \mathcal{M}_{T_{\text{Tycho}}}.$$

Hanson's thesis that observation is structured by conceptual frameworks is thus recast as: observation is shaped by inference under theory-constrained generative models, including their priors, likelihood

assumptions, and precision regimes [11, 16].

4.2 Seeing-as as Generative Inference

The core of Hanson's account of observation is not mere sensory registration but *seeing-as*: taking something *as* something under a conceptual organization [16]. Ambiguous figures illustrate that a fixed proximal stimulus can support distinct organizations (distinct "ascriptions").

Within PP, a natural computational analogue of seeing-as is *generative inference*. Given a generative model M encoding assumptions about latent causes z and their sensory consequences s, observation corresponds (approximately) to inferring

$$p(z \mid s, M),$$

that is, which latent causes best explain the sensory input under the model. In this sense, Gregory's dictum that "perception is a hypothesis" can be read as the claim that perceptual content reflects posterior inference rather than direct copying of stimulus structure [15, 23, 25, 26].

A key refinement, consistent with the bridge criteria and the PP discussion in Section 3, is that posterior inference is modulated by uncertainty. Precision (estimated reliability) controls the balance between top-down expectations and bottom-up mismatch signals [3, 10, 11]. Consequently, even for the same s, observers (or the same observer across contexts) can occupy different posterior regimes:

$$\text{Same } s \;\Rightarrow\; \text{different } p(z \mid s, M) \text{ under different } M, \Pi.$$

Here Π is schematic: in standard PP formulations, precision control is level- and channel-dependent, and it modulates which mismatch signals dominate updating rather than simply scaling a single error term.

Computationally, "seeing as X" can be idealized as occupying a posterior state concentrated around the X-region of latent space (or, more generally, as selecting one mode of a multimodal posterior), while perceptual reorganization corresponds to a shift in posterior mass across modes.

This provides a bridge-friendly structural correspondence: Hanson's personal-level seeing-as describes the role that PP models at the sub-personal level as posterior-guided inference under a generative model, with the strength and locus of theory-ladenness varying systematically with precision and hierarchical level (D1) [3, 10, 16].

4.3 Observation Statements as Report Policies over Posteriors

Scope note (perception vs. report). The present account does not require a strong thesis of cognitive penetration of early vision. Our target is *scientific observation* as a composite practice that typically involves (i) hierarchical perceptual inference (seeing-as as posterior-guided estimation under a generative model), (ii) precision allocation (which determines how "hard" or "soft" the influence of high-level expectations can be), and (iii) report policies that transform posterior states into public observation statements under goals, contexts, and norms. Accordingly, even in cases where early perceptual content is relatively encapsulated, substantial theory-ladenness can arise at higher inferential levels and in the policy-governed transition from posterior states to publicly evaluable reports [9]. This division of labor follows the three-layer target stated in the Introduction, which separates early input-processing, higher-level seeing-as (generative inference), and norm-governed public reporting.

As discussed in Section 2, observation statements are not neutral reports of raw facts but linguistically and conceptually articulated claims within a practice [16, 32]. A natural computational idealization is therefore not that observation statements are probability-to-text readouts, but that they are outputs of a *report policy* that selects what to say (and how strongly to say it) on the basis of an agent's posterior state, given goals and context.

Let u range over report options (e.g., categorical labels, hedged claims, upper bounds, or withholding judgment), and let G and C denote goal and context parameters. A minimal policy-based formulation is:

$$u^* \;=\; \arg\max_u \; \mathbb{E}_{z\sim p(z|s,M)}[\,U(u,z;\,G,C)\,]\,.$$

How to read U (epistemic scoring, not hedonic utility). The objective function $U(\cdot)$ is *not* a psychological or hedonic utility (comfort, persuasion, etc.). It schematizes a *publicly regulated epistemic scoring function*: it encodes which reports are licensed, how strongly they may be stated, and how trade-offs (informativeness vs. risk of false assertion) are handled under scientific reporting standards. Concretely, U can be read as encoding epistemic-scoring considerations such as accuracy and calibration (avoiding false assertion under uncertainty), informativeness (avoiding vacuous hedging), and community-imposed costs/standards for public commitments. This matters because observation statements are normative acts: they undertake commitments and license inferences within a shared practice, in the broadly Sellarsian sense that observation claims occupy positions in a space of reasons rather than merely expressing private states [32].

Making communal constraints explicit (N, I). Scientific reporting is constrained by community-level norms and measurement regimes. Let N denote reporting norms (admissible vocabulary, hedging conventions, confidence thresholds, error-reporting standards) and let I denote instrumentation/calibration regimes (accepted measurement models, correction procedures). Then the policy is more appropriately idealized as:

$$u^* \;=\; \arg\max_u \; \mathbb{E}_{z\sim p(z|s,M)}[\,U(u,z;\,G,C,N,I)\,]\,,$$

which links the computational state (posterior structure) to the public observation statement via explicit communal constraints [2, 12]. Here N should be read as constraining admissible report options and public scoring standards, while I fixes the instance-level calibration and measurement pipeline used to instantiate licensed likelihood families. For explicit boundary rules separating $\mathcal{L}_T$ from I and N from G, C, see Section 1.6.

Several familiar special cases illustrate how such policies can behave:

- **MAP-style reporting:** report the most probable label under the posterior.
- **Thresholded/hedged reporting:** report X only if posterior support exceeds a threshold; otherwise hedge or withhold.
- **Goal-sensitive reporting:** the same posterior can yield different verbalizations depending on whether the goal is archiving, warning, or hypothesis testing.

Toy report-policy class (detection-style). To make the policy form more concrete, let report options be $u \in \{\mathsf{detect}, \mathsf{evidence}, \mathsf{upper\text{-}bound}, \mathsf{withhold}\}$, where (roughly) detect is a categorical detection

claim, evidence is a hedged claim, upper-bound reports a constraint without detection, and withhold declines commitment. A minimal epistemic scoring function $U(u, z; G, C, N, I)$ can then include:

- **Error-cost trade-offs:** asymmetric costs for false positives vs. false negatives (stakes-sensitive via G, C).
- **Calibration / overclaim penalties:** penalties for reporting beyond what is warranted by calibration and uncertainty quantification (linked to I).
- **Informativeness:** rewards for informative commitments and penalties for vacuous hedging when stronger claims are warranted.
- **Norm penalties:** penalties for violating community thresholds, required uncertainty reporting, or admissible vocabulary (encoded by N).

Here N fixes which report forms and thresholds are permissible (e.g., "5σ"-style criteria, mandatory error bars, and licensed vocabulary for reporting), while I fixes the concrete preprocessing and calibration pipeline (baseline corrections, instrument settings, and parameterizations) used to instantiate licensed likelihood families in practice (cf. Section 1.6).

Why this matters for theory-ladenness. Because observers with different admissible model spaces (or different priors, likelihood assumptions, and precision regimes) can form different posteriors in response to the same s, they may also produce systematically different observation statements even when the sensory input is held fixed. Moreover, even holding posterior structure fixed, differences in G, C, N, I can induce report-level variation (D3) without any commitment to strong early-vision penetration. This preserves Hanson's emphasis on linguistic practice while keeping the computational bridge modest: observation statements are modeled as goal-, context-, and norm-sensitive extractions from posterior states, not as transparent readouts of theory-free data [9, 16].

4.4 Facts and Data: Two-Tier Stability and Model-Dependence

Hanson claimed that a "fact" is never a piece of naked data but always a *conditioned* item, stabilized within a background framework of description [16]. Historical and philosophical work on scientific practice also emphasizes that stabilized "facts" are not identical with raw data records, but are extracted and secured through methodological and instrumental practices [2, 6, 24].

To connect this to the bridge framework, it is useful to distinguish (i) data records from (ii) stabilized report classes. In the present notation, s can be treated as a data-bearing sensory or instrumental input, while a "fact" in scientific practice corresponds more naturally to a *publicly evaluable report type* u that is stabilized across contexts under shared constraints.

Computationally, posterior structure is still central: report eligibility and stability depend on what is supported by $p(z \mid s, M)$ under an admissible model. But it is a mistake to identify facthood with posterior structure alone. Instead, one can idealize scientific facthood as a two-tier stability condition:

- **Posterior support (intra-agent / intra-model):** a report type u is licensed by an agent's posterior state under a model M (and relevant policy constraints).

- **Reproducible stability (inter-agent / inter-lab):** the same report type is robustly reproduced across agents/labs under shared calibration regimes and reporting norms (I, N), and remains stable under reasonable variations of measurement/noise assumptions (including explicit model checking) [2, 12].

Methodologically, the same toolbox that secures fact-stability also supports anomaly diagnosis: posterior predictive checks and residual diagnostics help distinguish noise-like deviations from persistent structured misfit that motivates model-class or constraint revision.

This framing preserves Hanson's insight that facts depend on background frameworks while also explaining how objectivity is achieved in practice: not by theory-free givens, but by stabilized procedures for extracting phenomena from data and for regulating public reporting. On this view, "fact" is neither a private sensation nor a mere posterior state; it is a stabilized public commitment supported by posterior evidence under communal constraints.

Finally, this is exactly where anomalies become operational. An "anomaly" in the strong sense is not a single surprising datum but a *persistent, structured residual* that survives best within-model tuning and plausible measurement-model repairs, thereby motivating a revision of admissibility constraints (D2) [2, 12, 26]. This prepares the ground for the multi-tier dynamics developed below.

4.5 The Construction of Causality as Structure in Generative Models

As discussed in Section 2, Hanson treated causality not as a property transparently read off from event streams, but as a selective organization of events into explanatory relations under a theoretical standpoint [16]. Which variables count as causes, which links are salient, and which counterfactual contrasts matter depends on background commitments and explanatory aims.

Within PP and related Bayesian modeling frameworks, directed dependence structure is typically encoded in the generative model. A schematic hierarchical dependence can be written as

$$z_L \to z_{L-1} \to \cdots \to z_1 \to s,$$

where higher-level latent states generate lower-level states and ultimately sensory input. Hierarchical Bayesian models of perception provide concrete instantiations of such generative structure [23, 25].

Generative dependence vs. interventionist causation. It is crucial to keep distinct two notions of "causal structure." Directed dependencies in a generative model encode a *model-relative* asymmetry of generation and conditional dependence; they do not automatically coincide with interventionist or manipulability notions of causation. Interventionist frameworks (e.g., causal Bayesian networks and structural equation models) impose additional constraints connecting directed structure to counterfactual and interventional claims [29, 40]. In this paper, "causal structure" is used primarily in the generative-model sense appropriate for PP-style explanation: it characterizes how the model organizes explanatory dependence and prediction. Bridging this to interventionist causation is an important further project, but it is not required for the present bridge claim about theory-ladenness and the organization of seeing-as and explanation.

On the present reading, causal explanation corresponds to selectively articulating parts of the generative structure under a theory-conditioned model space (Section 1.5): to explain is to locate an event within

a model's dependency organization and to identify which latent variables and links do the explanatory work. This mirrors Hanson's claim that causal chains are not simply discovered but constructed under a theoretical stance [16].

4.6 Abduction as Hypothesis Formation, Model Search, and Model Selection

Hanson understood scientific discovery not as passive accumulation of data but as active hypothesis formation and pattern detection [16]. His reliance on Peirce's notion of abduction is well known [30], and later epistemology connects abduction to inference to the best explanation (IBE) [17].

Within a Bayesian vocabulary, abduction can be *partly* reconstructed as model comparison and selection: competing models $M_1, M_2, \ldots$ are compared by their posterior credibility,

$$p(M_i \mid s) \propto p(s \mid M_i)\, p(M_i),$$

where $p(s \mid M_i)$ is model evidence and $p(M_i)$ is a prior over models [12, 26]. In a simplified setting, one may write

$$M^* = \arg\max_{M_i} p(M_i \mid s).$$

However, abduction is not merely *selection among a fixed list*. A central feature of discovery is that the space of admissible hypotheses itself can change. In the present framework, this corresponds to changes in theory-level constraints, i.e. revisions of $\mathcal{M}_T$ (Section 1.5). Accordingly, the most faithful bridge-friendly idealization treats discovery as a *three-tier* process:

1. **Within-model updating:** hold M fixed and update parameters or latent-state estimates so as to reduce mismatch (C1; within-model dynamics).
2. **Between-model selection:** compare alternative models within a fixed admissible space $M \in \mathcal{M}_T$ (C2; model comparison).
3. **Model-space (constraint) revision:** revise theory-level admissibility constraints $T \mapsto T'$ so that $\mathcal{M}_T$ changes (C3; reorganization of what is admissible as explanation and observation).

This explicit separation is crucial for respecting Hanson's discovery structure while preserving the bridge criteria. It also blocks a common objection that Bayesian updating can only describe gradual change: the framework allows both continuous adjustment within a model, discrete shifts across models, and (when required) reorganizations of the admissible model space itself [24, 26].

Finally, none of this identifies Bayesian model choice with normative scientific rationality. Although evidence and complexity trade-offs parallel familiar epistemic virtues, they do not by themselves yield normative conclusions. The present use of Bayesian/PP structure is explanatory and bridge-oriented, not a reduction of normativity [18, 40].

4.7 Formulating Hanson's Theory-Ladenness in Predictive-Processing Terms

On the basis of the correspondences developed above, Hanson's theory-ladenness can be expressed using a unified PP vocabulary:

Theory	Meta-level constraint on admissible model space $\mathcal{M}_T$ (variables, structure, likelihood/noise, durable precision expectations)
Seeing-as	Posterior-guided inference over latent causes, $p(z \mid s, M)$
Observation statement	Report policy over posteriors (goal-, context-, norm-, and instrument-sensitive extraction from $p(z \mid s, M)$)
Fact	Stabilized report type supported by posterior evidence; scientific facthood requires reproducible stability under shared N, I (data/phenomena extraction)
Causal explanation	Selective articulation of model-relative dependency/causal structure encoded in the generative model
Hypothesis formation (abduction)	Search/selection with multi-tier dynamics (within-model updating; between-model selection) and, when required, constraint revision ($T \mapsto T'$)

These correspondences are intended as more than loose metaphors: they specify role-to-role mappings that can be made explicit in Bayesian terms. As a simplified illustration, one possible report policy is MAP-style labeling,

$$u_{\text{MAP}} \;=\; \arg\max_{z} p(z \mid s, M),$$

but this is only one policy among many. Thresholded and goal-sensitive reporting, as well as communal constraints (N, I), are captured by the decision-theoretic formulation in Section 4.3.

To emphasize again: this mapping does not identify the personal level ("Kepler reasoned in this way") with the sub-personal level ("the brain computed in this way"). They may be causally related, but they remain distinct explanatory levels. The claim here is structural correspondence (a realization relation), not reduction [5, 22].

4.8 Summary: Validity and Limits of the Computational Reinterpretation

This section concludes by summarizing both the achievements and the limitations of the computational reconstruction.

Achievements: making the correspondence explicit. We reconstructed Hanson's core roles—observation/seeing-as, theory, causal explanation, and hypothesis formation—in a computational vocabulary that distinguishes generative inference, precision-weighted mismatch, report policies, and multi-level revision dynamics. This clarifies how Hanson's insights can be rendered as a structured bridge between personal-level patterns and sub-personal computational roles, while avoiding the slogan *theory = prior* by treating theories as constraints on admissible model space [2, 12, 29, 40]. It also sharpens the locus-of-variation story: theory-ladenness is expected to vary with precision and hierarchical level (D1), and anomalies are operationalized as persistent structured residuals that motivate revision (D2) [3, 10, 12].

Limitations: idealizations and open gaps. Several limitations remain:

- PP/FEP is a family of modeling strategies rather than a settled neurobiological theory; the present bridge relies only on weak commitments (Section 3.8) [4, 11].

- The link between theory-level admissibility constraints and learned conceptual structures remains incomplete; development, training, and social organization likely matter essentially for how T is acquired and revised [37].
- The mapping from posterior states to natural-language observation statements is modeled only schematically via report policies; a fuller account would require further work in pragmatics and philosophy of language (cf. Sellars on inferential role) [32].
- Generative dependence in PP does not automatically yield interventionist causation; connecting model-relative causal structure to interventionist explanation remains a further project [29, 40].
- Variational/evidence-based interpretations parallel certain epistemic virtues but do not by themselves ground normative claims about scientific rationality [12, 18].

These limitations do not undercut the main contribution of this section. They clarify the status of the proposal: a computationally sharpened, failure-sensitive bridge hypothesis that preserves Hanson's personal-level structure while opening a route to explicit modeling.

In the next section (Section 5), we examine how far this framework can go in reconstructing concrete episodes of scientific discovery, and we assess the prospects and limits of a unified model spanning philosophy of science, history of science, and cognitive science.

5 A Computational Model of Scientific Discovery

Section 4 provided a computational reconstruction of the theory-ladenness of observation, articulating a bridge between Hanson's personal-level roles (seeing-as, observation reports, anomaly, discovery) and PP/Bayesian sub-personal roles (generative inference, precision-weighted mismatch, report policies, and multi-level revision). The aim of the present section is to extend that reconstruction from the comparatively static thesis of theory-laden observation to its dynamic counterpart: a computational model of *discovery* and *theory change*.

The guiding thought is that discovery is not an external add-on to theory-ladenness. Rather, if observation is inference under theory-conditioned admissibility constraints, then discovery is (sometimes) the revision of those constraints in response to persistent structured misfit. This section therefore develops a bridge-friendly account of the *theory-ladenness of discovery*: how anomalies can drive within-model retuning, between-model selection, and, when required, revisions of the admissible model space itself (cf. C1–C3 in Section 1.4).

In Marr's and Dennett's sense, the present account is a bridge hypothesis: it specifies a role- and constraint-preserving structural correspondence between personal-level episodes of discovery and sub-personal inferential dynamics, without identifying the levels [5, 7, 22, 27].

Scope note (individual vs. community). Predictive-processing models primarily target sub-personal mechanisms in individual agents, whereas scientific discovery and revolution are also embedded in community-level institutions, instruments, and norms. Accordingly, we develop the bridge mainly at the level of individual inferential organization (personal/sub-personal), while treating communal stabilization as constraint-bearing inputs rather than as something reduced to neural computation. In the bridge

framework, communal constraints enter explicitly via N (reporting norms) and I (instrumentation/calibration regimes), which shape the policy-governed transition from posterior states to publicly evaluable observation reports and the stabilization of report types as "facts" [2, 12].

5.1 Positioning Hanson's Account of Discovery: Observation → Anomaly → Discovery

In *Patterns of Discovery*, Hanson argues that scientific discovery is not a simple extension of data accumulation. Rather, it involves a qualitative shift in which the observer comes to see the world differently, often prompted by an anomaly that cannot be coherently integrated within the current theoretical organization [16]. In this respect, Hanson rejects the idea that discovery can be cleanly separated from the epistemic life of observation, a point reinforced by later historical work on scientific practice [6, Ch. 1].

Hanson's schematic pattern can be summarized as:

1. **Observation**: seeing the world under an existing theory (a given seeing-as)
2. **Anomaly**: a persistent mismatch relative to the expected pattern
3. **Discovery**: acquisition of a new way of seeing that renders the mismatch intelligible

For Hanson, an anomaly is not merely an irregular datum. Its significance lies in the fact that it resists integration within the operative explanatory scheme. This is why anomalies must be distinguished from mere noise, measurement error, or one-off outliers.

In predictive-processing terms, the closest analogue to the anomaly role is not a bare prediction error at a single time point but a *regime* of precision-weighted mismatch that persists and exhibits structured residual patterning (cf. discriminator D2) [10, 11, 12]. For expository purposes one may write

$$\varepsilon_t \approx \Pi_t \, (s_t - \hat{s}_t),$$

but this is a didactic idealization; in standard PP formulations, mismatch signals are distributed across hierarchical levels and their influence depends on precision assignment (Section 3.4).

An anomaly, in the bridge sense, corresponds to a situation in which mismatch signals remain persistently salient *after* reasonable within-model responses have been attempted (parameter retuning, measurement-model repair, and precision reallocation; see Section 5.2 below). When such structured misfit survives, revision pressure shifts from local tuning to model (or model-space) change, matching the Hanson pattern: observation → anomaly → discovery.

Operational gloss on "structured misfit." In what follows, "structured residuals" refers to misfit patterns that are stable across best-fitting parameter retuning and plausible likelihood/measurement variants, and that remain visible under standard diagnostics (e.g., posterior predictive checks and residual analyses), rather than dissipating as noise under reasonable re-specification. For a short practice-oriented checklist of diagnostics (posterior predictive checks, residual structure, and robustness to noise-model variants), see discriminator **D2** in Section 1.12.

The PP cycle of inference → mismatch → revision thus offers a formal expression of Hanson's dynamic sequence. This remains a structural correspondence between sub-personal inferential dynamics

and personal-level episodes of surprise, discrepancy, and reinterpretation, not an identity claim. The bridge target is shared abstract organization, not reduction.

5.2 A Taxonomy of Anomaly Resolution: Parameter, Measurement, Precision, and Constraint Revision

To avoid conflating distinct responses to mismatch, we distinguish three "within-model" routes that can often resolve apparent anomalies without requiring genuine between-model change, before turning to cases that plausibly require revising admissibility constraints.

Within-model(A): parameter/prior retuning (fixed structure and measurement model). Here the model class M is held fixed while parameters (or prior settings within the class) are adjusted to reduce misfit. This corresponds to routine learning or retuning within a stable explanatory framework (normal-science-like adjustment) [12, 26]. (In schematic terms: θ-level retuning within a fixed M and fixed $\mathcal{L}_T/I$.)

Within-model(B): measurement/noise-model repair (instrument-model revision). Here the target of revision is not the world-model structure but the likelihood/measurement assumptions, including calibration regimes I and admissible likelihood families $\mathcal{L}_T$. Many dramatic discrepancies dissolve once the measurement model is repaired or re-specified (cf. the data/phenomena distinction) [2, 12]. This route is central for diagnosing when "anomaly" is an artifact of the instrument-model interface rather than of the world-model. (In schematic terms: revisions to $\mathcal{L}_T$ and/or I with the world-model structure held fixed.)

Within-model(C): precision reallocation and attentional reweighting. Even with fixed structural assumptions, redistributing precision across levels/channels can change which mismatch signals dominate updating. This route is especially relevant to cases where the personal-level episode looks "Hansonian" (surprise and reinterpretation) but no model-class revision is computationally required (failure mode **F1**) [3, 9, 10]. In bridge terms, such cases can generate report-level and interpretation-level variation without reorganizing admissible model space. (In schematic terms: reallocation of Π across levels/channels; distinct from report-policy variation in G, C, N (D3).)

Between-model and beyond: revision of model class and admissibility constraints. When structured residuals persist across best-fitting parameters, plausible measurement/noise-model repairs, and reasonable precision reallocations—and especially when the relevant signals are high-precision and robustly replicated—the misfit plausibly targets the admissibility constraints themselves (e.g., $\mathcal{V}_T, \mathcal{S}_T$), motivating model-class change and reorganization of what counts as an admissible observation, fact, or explanation (criterion **C3**) [12, 24, 26]. This is the bridge-theoretic sense in which an anomaly can be revision-driving rather than merely inconvenient.

5.3 Reanalyzing Kepler's Discovery of Elliptical Orbits

One of Hanson's central historical illustrations is Kepler's discovery of elliptical planetary orbits [16]. Here we offer a structural reanalysis through the framework of PP and Bayesian inference. The aim is not

to reproduce full historical detail, but—drawing on standard scholarship in the history of early modern science [6, 38]—to make explicit the inferential roles at issue in a way that is bridge-friendly.

Throughout, the Bayesian/PP vocabulary is used as an idealizing representational framework for the *structure* of Kepler's reasoning, not as a claim that Kepler literally performed probabilistic calculations.

A strong admissibility constraint: uniform circular motion

By the late sixteenth century, a widely entrenched methodological commitment in European astronomy favored uniform circular motion: planetary trajectories were expected to be composed of circles (often implemented via deferents and epicycles), and departures from circularity were typically handled through auxiliaries rather than by revising the basic orbit-shape assumption [6, 38]. In the notation of Section 1.5, this functions as a theory-level admissibility constraint restricting the relevant region of model space.

Historically, this commitment functioned less like a soft preference and more like a methodological constraint on what counted as an admissible basic explanation of planetary motion. The ideal of saving the phenomena by uniform circular motions was embedded in a package of mathematical intelligibility conditions, representational habits, and standards of astronomical acceptability: departures from circularity were typically treated as something to be accommodated by auxiliaries (epicycles, eccentrics, equants, coordinate choices, and related devices) rather than as a reason to treat non-circularity as an eligible primitive orbit-shape hypothesis [6, 38].

In bridge terms, this matters because it supports the intended diagnosis of an *eligibility regime*. Even if one can retrospectively parametrize "ellipse" as "circle plus an eccentricity parameter," the historically operative constraint was that $e \neq 0$ was not treated as a live explanatory degree of freedom at the level of basic orbit shape. Accordingly, what changes in the Kepler episode is not merely the estimate of a parameter already treated as admissible, but the admissibility status of a structural option: the model space itself is reorganized so that non-circularity becomes an acceptable basic explanatory form [6, 38].

Schematically, one may represent the restriction as privileging a model regime in which orbit-shape hypotheses are effectively constrained to circular constructions. Lower-level parameters (epicycles, eccentrics, equants, etc.) are then adjusted to fit observational data. The critical point for the bridge is that "circle" here functions not merely as a parameter value but as an eligibility constraint embedded in a broader framework of modeling and interpretation.

Persistent structured misfit: the Martian anomaly

Highly precise observational constraints (notably those associated with Mars) resisted accommodation by standard circular–epicyclic resources [6, 38]. In bridge terms, this is a case where mismatch is not a one-off deviation but a persistent structured residual pattern.

Computationally, such a case corresponds to a regime in which mismatch signals remain salient under high effective precision and survive routine within-model responses (A) parameter retuning and (B) plausible measurement-model repair. The intended diagnosis is not simply "large error," but *structured* misfit robust to best within-model tuning (D2) [12, 26].

Model search under shifting loci of error

Kepler's extended period of trial and correction can be reinterpreted as model search under uncertainty about the locus of misfit:

- refining parameters within a circle-privileging class M_{circle},
- exploring eccentric-circle and epicyclic variants within that class,
- reconsidering whether the admissibility constraints themselves should be revised.

Episodes in which Kepler discovered and corrected computational mistakes illustrate the within-model reassignment of error loci (procedure vs. data vs. measurement assumptions), while the eventual move to ellipses illustrates a shift in admissibility constraints [38].

A potential objection (nested models) and the intended reply. A statistical worry is that "circle vs. ellipse" looks like a nested-model case: a circle is a limiting case of an ellipse (e.g., $e = 0$). Our point, however, is not a claim about what Kepler explicitly parameterized. Rather, the historically entrenched commitment functioned as a methodological *eligibility regime*: non-circularity was treated as ineligible as a basic explanatory form, with departures managed via auxiliaries. Representing this as a hard restriction on admissible model space (Section 1.5) captures the bridge-relevant structure: relaxing the eligibility constraint so that $e \neq 0$ becomes admissible constitutes a revision of admissibility constraints (a change in $\mathcal{M}_T$), not merely estimation of a parameter value within a fixed regime [6, 38] (cf. Appendix A for a minimal nested-model formalization).

Table 2: Constraint-change diagnosis in the Kepler case (schematic).

Component	**Before (circle-privileging regime)**	**After (ellipse-admissible regime)**
$\mathcal{V}_T$ (variables)	Orbit-shape hypotheses effectively restricted to circular constructions; deviations treated as auxiliaries	Orbit-shape hypotheses expanded so that eccentricity becomes an admissible explanatory variable
$\mathcal{S}_T$ (structure)	Admissibility constraint strongly favors $e = 0$ as the only eligible basic structure; "ellipse" not treated as eligible geometry	Constraint relaxed: $e \neq 0$ becomes eligible; structure permits focal geometry and offset relations
$\mathcal{L}_T$ (likelihood/measurement)	Misfit partly interpretable as noise or procedural/instrumental artifacts under favored circular constructions	Measurement assumptions tightened and checked; persistent residuals treated as model-relevant structure rather than dismissible noise
Π_T (precision expectations)	Tolerance for discrepancy can keep misfit within auxiliary repair; key streams may be treated as less diagnostically decisive	High precision assigned to key constraints makes structured residuals salient and revision-driving

The point of this diagnosis is not that Kepler explicitly manipulated these components, but that the historical shift can be represented as a change in admissibility constraints: the transition is from treating $e = 0$ as effectively mandatory to treating e as a live explanatory degree of freedom. This is why the shift is "between-model" in the methodological sense relevant to theory-ladenness: it changes what counts as an eligible explanation and thereby reorganizes observation and facthood.

Adopting the ellipse: model comparison under revised admissibility constraints

Kepler ultimately adopted the model according to which planetary orbits are ellipses with non-zero eccentricity ($e \neq 0$). In Bayesian notation, this can be represented schematically as selecting the model with highest posterior credibility within the admissible space:

$$M^* = \arg\max_{M \in \mathcal{M}_T} p(M \mid s) \quad \propto \quad p(s \mid M)\, p(M),$$

where $p(s \mid M)$ is model evidence (marginal likelihood) and $p(M)$ is a prior over model classes [12, 26].

Despite strong antecedent preference for circular constructions, the ellipse achieved sufficiently greater evidential support (and explanatory coherence) to shift posterior mass toward an ellipse-admissible regime. In Hanson's terms, Kepler acquired a new way of seeing: Mars was no longer seen as a perturbed circle but as an ellipse [16]. In bridge terms, this corresponds to revising theory-level admissibility constraints and thereby reorganizing the space of eligible explanations.

Kepler's case thus illustrates how persistent structured misfit can drive not only parameter adjustment but also model-space revision. It also highlights that discovery is not a one-directional sequence ("data $\rightarrow$ theory") but a feedback process in which expectations guide interpretation while anomaly diagnoses motivate restructuring [6, 38].

5.4 Reinterpreting Scientific Abduction through Bayesian Inference

Hanson's account of discovery is grounded in Peircean abduction—hypothesis formation in response to surprising facts [16, 30]. Harman's influential notion of inference to the best explanation (IBE) can be seen as a modern restatement of the abductive pattern [17].

Within Bayesian inference, the *selection* aspect of abduction can be made explicit:

$$p(M \mid s) \propto p(s \mid M)\, p(M).$$

To account for anomalous data s, one may provisionally adopt a model M that achieves high posterior credibility, balancing evidential fit (model evidence) against prior credibility and complexity trade-offs [12, 21, 26]. In an idealization, the "best explanation" corresponds to a model maximizing $p(M \mid s)$.

However, abduction is not merely selection among a fixed list. A distinctive feature of discovery is *model search and model construction*: the space of admissible hypotheses can change (Section 1.5). Accordingly, a bridge-friendly reinterpretation treats abduction as a multi-stage process that includes:

1. **model-space articulation:** specifying (or revising) what is admissible as a candidate model class $\mathcal{M}_T$;
2. **within-space comparison:** evaluating candidates by their evidential support and complexity trade-offs;
3. **selection and stabilization:** adopting a model and stabilizing associated report practices under communal constraints (N, I).

This preserves the spirit of abduction while avoiding the over-strong claim that abduction *is* Bayesian

model selection. Bayesian structure captures a core *constraint pattern* (fit, complexity, and updating), while leaving room for the creative and socially mediated aspects of hypothesis generation.

5.5 Scientific Revolutions as Revisions of High-Level Constraints (Paradigms)

Kuhn's notion of a paradigm shift can be reinterpreted, within the present bridge framework, as a revision of high-level admissibility constraints and thus an alteration of model space [24]. Normal science corresponds to within-model and within-space adjustments under stable constraints; revolutions correspond to changes in the constraint package.

In the notation of Section 1.5, ordinary change can often be represented as within-model updating:

$$\theta \leftarrow \theta' \quad \text{with fixed } M \in \mathcal{M}_T,$$

and, more broadly, as selection among models within a fixed admissible space $M \in \mathcal{M}_T$.

By contrast, a revolution involves revising the admissibility constraints:

$$T \longrightarrow T', \qquad \text{hence} \qquad \mathcal{M}_T \longrightarrow \mathcal{M}_{T'}.$$

When T changes, the space of eligible explanations changes, and with it the organization of seeing-as, observation reporting, and explanatory practice. Kuhn described this as the world looking different; Hanson analyzed it as a change in seeing-as [16, 24].

From a PP/FEP perspective, one may also describe such shifts as responses to persistent system-level misfit: when mismatch cannot be reduced by within-model tuning and plausible measurement-model repairs, more radical revision becomes pressing [11, 12]. This does not reduce scientific rationality to free-energy minimization; it offers a computational vocabulary for the role structure (persistent misfit $\rightarrow$ constraint revision) that parallels the historical pattern of revolutionary change.

5.6 Connecting Scientific Model Revision with Contemporary AI and Machine Learning

The aim of this brief connection is purely structural: it highlights that the same fit/complexity and misfit-diagnosis patterns arise in modern model-building practice, without suggesting a reductive identification of scientific theories with machine-learning models.

Many approaches in contemporary AI and machine learning can be characterized as constructing models that predict and explain data. Even when methods are not explicitly Bayesian, their dynamics often exhibit structural similarities to Bayesian learning and model selection (fit vs. complexity, model class choice, and adaptation under distribution shift) [1, 14, 26, 28].

Relevant structural parallels include:

- **Parameter learning within a fixed model class** (within-model updating);
- **Architecture and hyperparameter search** (between-model comparison/selection);
- **Regularization as complexity control** (often interpretable in Bayesian terms);
- **Train–test mismatch / dataset shift** (a practical analogue of anomaly diagnosis under misfit).

These parallels should be treated cautiously. They are structural, not reductive: scientific theories carry semantic and normative dimensions, and their evaluation is community-regulated and instrument-mediated. Still, Bayesian generative modeling provides a useful abstract framework for capturing some of the inferential role structure that underwrites both machine learning practice and scientific reasoning [26, 37].

5.7 Summary: From the Theory-Ladenness of Observation to the Theory-Ladenness of Discovery

The theory-ladenness of observation is not merely a static feature of perceptual organization; it deepens in the dynamics of discovery. If observation is inference under theory-conditioned admissibility constraints, then discovery is (sometimes) the revision of those constraints in response to persistent structured misfit.

In bridge terms, the key transitions can be summarized as follows:

- Hanson's pattern *observation* $\rightarrow$ *anomaly* $\rightarrow$ *discovery* corresponds to inference under a generative model $\Rightarrow$ persistent precision-weighted structured misfit $\Rightarrow$ within-model updating, between-model selection, and (when required) constraint revision $T \mapsto T'$ [11, 12, 16].
- The anomaly taxonomy (Section 5.2) distinguishes cases resolved by parameter retuning, measurement/noise-model repair, or precision reallocation from cases that plausibly require revising admissibility constraints (C3; D2) [2, 10, 12].
- Kepler's case illustrates how a historically entrenched eligibility regime can be forced open by persistent structured residuals, motivating a shift to an ellipse-admissible model space and reorganizing what is seen and reported as a fact [16, 38].
- Abduction and IBE can be partially reconstructed as model comparison under evidence and priors, while preserving the fact that discovery often involves revising the admissible hypothesis space itself [17, 26, 30].
- Kuhnian revolutions can be represented as revisions of high-level constraint packages (and sometimes of associated norms/instruments), i.e. $(T, N, I) \mapsto (T', N', I')$, rather than as mere parameter updates [2, 24].

Thus, this section has provided a computationally sharpened, failure-sensitive bridge for the dynamics of discovery. Discovery is not a matter of "pure data" forcing a new theory. It is a process in which misfit is diagnosed under a theory-conditioned model space, within-model repairs are attempted, and only persistent structured residuals motivate more radical revisions of admissibility constraints. In this way, predictive-processing and Bayesian modeling supply a naturalized role vocabulary for discovery without reducing normative epistemology or community-regulated practice to sub-personal computation [5, 22].

Section 6 evaluates the contributions and limitations of this approach and outlines prospects for integrated research across philosophy, neuroscience, and computational modeling.

6 Conclusion

This study has sought to reformulate the classical philosophical problem of the theory-ladenness of observation, as articulated by Hanson, within the contemporary framework of predictive processing (PP)

and Bayesian modeling in cognitive science.

Status of the bridge (what it is and is not). The bridge hypothesis defended in this paper is *neither* a claim of mechanistic reduction (*personal* $\rightarrow$ *sub-personal* identity) *nor* an attempt at epistemic justification (deriving norms of scientific rationality from PP/FEP). Its intended output is more modest and more specific: it states *level-coherent constraints*—role- and constraint-preserving conditions under which personal-level patterns (observation–anomaly–discovery, theory-ladenness, reporting practices) can be *jointly modeled* by sub-personal computational roles (hierarchical generative inference, precision-weighted mismatch, and multi-level revision). Accordingly, if these constraints fail, the correspondence fails; this is why we stated explicit success conditions and failure modes (Section 1.4) and operational discriminators (Section 1.12).

Section 2 reconstructed Hanson's analysis of theory-ladenness, emphasizing that observation has the structure of seeing-as and depends essentially on concepts, language, and background theories [16, 32]. Section 3 surveyed PP-style computational theories of perception, outlining how perception and learning can be modeled as approximate inference under hierarchical generative models with precision-weighted mismatch signals [3, 4, 10, 11, 20, 34].

Section 4 made the bridge explicit: seeing-as was modeled as posterior-guided inference under a generative model; theories were treated not as numerical priors but as meta-level constraints on admissible model space (Section 1.5); observation reports were modeled as goal-, context-, and norm-sensitive *report policies* over posterior states (Section 4.3); and "facts" were connected to stabilized report types supported by posterior evidence, with a distinction between intra-agent stability and inter-agent reproducibility under shared instrumentation/calibration and reporting norms (I, N) [2, 12, 26].

Section 5 extended this structure to the dynamics of discovery, articulating (i) a taxonomy of anomaly resolution (parameter retuning, measurement-model repair, precision reallocation) and (ii) conditions under which persistent structured misfit motivates between-model selection and, when required, admissibility-constraint revision $T \mapsto T'$ (Section 5.2) [12, 16, 17, 24, 26, 30].

Importantly, the slogan "theory = prior" is *not* endorsed. Theories are semantic and normative structures; they are not literal probability distributions. The intended claim is instead: insofar as theories and paradigms shape observation at the personal level, one may model their sub-personal realization as constraints on admissible model spaces (including admissible prior families, likelihood/measurement assumptions, and durable precision expectations) [18, 29, 40]. Accordingly, theory change is best decomposed into (i) within-model updating, (ii) between-model selection within a fixed admissible space, and (iii) when required, revision of admissibility constraints $T \mapsto T'$ (hence $\mathcal{M}_T \mapsto \mathcal{M}_{T'}$).

Scope note (individual vs. community). Predictive-processing models primarily target individual sub-personal mechanisms; scientific objectivity and stabilization also depend on community-level institutions, instruments, and norms. In the present bridge framework, communal constraints are not reduced to neural computation but enter explicitly as constraint-bearing inputs via N (reporting norms) and I (instrumentation/calibration regimes), which shape the policy-governed transition from posterior states to publicly evaluable observation reports and the stabilization of report types as "facts" [2, 12].

Explanatory payoffs (recap). To highlight what the bridge adds beyond mere restatement (cf. Section 1.11), the present framework yields several modest but substantive payoffs:

Empirical and methodological touchpoints. The bridge is not merely interpretive: it yields operational touchpoints for where theory-ladenness should vary (D1), how strong anomalies can be diagnosed (D2), and how report-level variation can dissociate from perceptual content (D3) (Section 1.12).

- a principled locus-of-variation story via precision allocation (where theory-ladenness is likely to be "hard" vs. "soft") [3, 10];
- a failure-sensitive, multi-tier diagnostic for mismatch episodes: when discrepancies are resolvable by within-model routes (parameter retuning, measurement repair, precision reallocation) versus when they motivate between-model selection or constraint revision (Sections 1.4, 5.2) [12, 26];
- a minimal action-oriented model of observation statements as *report policies* over posterior states, explicitly sensitive to goals, contexts, and communal constraints (N, I), rather than probability-to-text readouts [2, 32].

6.1 Summary of Contributions

The main contributions of this study can be organized under three headings:

1. **A computational reinterpretation of the theory-ladenness of observation**

   Hanson's seeing-as was reconstructed as posterior-guided inference under a generative model. Theoretical background was modeled not as a single prior but as a meta-level constraint on admissible model space (variables, structures, likelihood/measurement assumptions, and durable precision expectations; Section 1.5). Observation reports were modeled as outputs of report policies over posterior states under goals, contexts, and communal constraints (G, C, N, I; Section 4.3). "Facts" were treated not as posterior states *as such*, but as stabilized report types supported by posterior evidence and secured via reproducibility under shared calibration and reporting regimes [2, 12, 16, 26].

2. **A predictive-processing reconstruction of anomaly roles in discovery**

   Scientific anomalies were reconstrued not as single surprises but as persistent, precision-weighted *structured misfit* regimes, diagnosed after reasonable within-model responses (parameter retuning, measurement/noise-model repair, and precision reallocation) have been attempted (Section 5.2) [11, 12]. Kepler's adoption of elliptical orbits was analyzed as a constraint-change episode in which a historically entrenched admissibility regime was forced open by persistent residual structure, motivating a shift to an ellipse-admissible model space [6, 16, 38].

3. **A unified computational account of theory change and revolutions as constraint revision**

   Kuhnian paradigm shifts and Hansonian transformations in seeing-as were modeled as revisions of admissibility constraints and reorganizations of the space of eligible models ($T \mapsto T'$, hence $\mathcal{M}_T \mapsto \mathcal{M}_{T'}$), often accompanied by shifts in instrumentation and reporting regimes ($(T, N, I) \mapsto (T', N', I')$) [2, 6, 24]. This analysis distinguished continuous within-model updating from discrete between-model transitions and, when required, model-space (constraint) revision within a single Bayesian role vocabulary [12, 26].

Taken together, these contributions integrate Hanson-style theory-ladenness, Peircean abduction, and Kuhnian theory change into a single, failure-sensitive bridge framework grounded in contemporary computational cognitive science.

6.2 An Integrated Understanding of Observation, Theory, and Perception

This study has argued that observation, theory, and perception need not be treated as independent layers. For the purposes of the bridge model, they can be fruitfully understood as roles within a hierarchy of generative inference: abstract background constraints shape admissible models; intermediate representational organization supports seeing-as and categorization; and lower-level sensory signals constrain inference via precision-weighted mismatch [3, 10, 11, 25, 31].

A key clarification is that this is a *structural correspondence*, not a one-to-one identity of personal-level theories with neural states. At the computational level, seeing-as can be idealized as posterior-guided inference under an admissible model:

$$p(z \mid s, M) \propto p(s \mid z, M)\, p(z \mid M), \qquad M \in \mathcal{M}_T,$$

where theory T constrains admissible model classes $\mathcal{M}_T$ rather than being identified with any single prior.

In Hanson's canonical example, Kepler and Tycho share the same proximal input s but, under different admissibility constraints, occupy different posterior regimes and hence different seeing-as:

$$\text{Same } s \quad \Rightarrow \quad \text{different } p(z \mid s, M) \text{ (under different } \mathcal{M}_T, \Pi).$$

Likewise, illusions and hallucinations can be understood as systematic outcomes of the balance among priors, likelihood assumptions, and precision weighting, suggesting that "pure observation" is unlikely to provide a theory-free baseline for scientific purposes, even when early perceptual processing is relatively constrained [15, 23, 35].

The bridge also clarifies why observation statements are not transparent readouts of posteriors. Observation reports are modeled as policy-governed public commitments, sensitive to goals and contexts and constrained by communal norms and measurement regimes:

$$u^* \;=\; \arg\max_{u} \; \mathbb{E}_{z \sim p(z \mid s, M)} \big[\, U(u, z;\, G, C, N, I) \,\big]\,,$$

capturing Hanson's and Sellars' insistence that observation claims belong to a norm-governed practice rather than to a myth of the given [2, 16, 32].

Finally, PP accounts of precision offer a principled way to reframe debates on cognitive penetration. High precision assigned to low-level features can yield robustness against top-down influence, helping to explain why some illusions persist even when one "knows better," while ambiguity and higher-level interpretation can remain more theory-sensitive. This yields a graded alternative to binary penetration/non-penetration claims [3, 9, 36].

6.3 Scientific Discovery and the Structure of Model Revision

This study has also argued that scientific discovery is not merely an accumulation of observations, but is characterized by recurrent cycles of model-based interpretation and revision. At the level of roles, the core cycle can be expressed as:

$$\text{Inference under } M \in \mathcal{M}_T \longrightarrow \text{mismatch diagnosis} \longrightarrow \text{revision (within/between/constraint).}$$

Kepler's case illustrates how discovery can be represented (as an idealization) in terms of Bayesian model comparison and selection under historically structured admissibility constraints [6, 12, 38]. Anomalies that trigger discovery correspond to persistent, precision-weighted structured residuals that survive best within-model tuning and plausible measurement-model repairs (D2) [2, 12].

Peircean abduction and IBE can be *partly* reconstructed within this vocabulary. Bayesian model comparison makes explicit one constraint pattern:

$$p(M \mid s) \propto p(s \mid M)\, p(M),$$

but discovery typically involves not only selection among a fixed list but also model search and, when required, admissibility-constraint revision $T \mapsto T'$ (hence $\mathcal{M}_T \mapsto \mathcal{M}_{T'}$) [17, 26, 30].

From this standpoint, scientific change can be summarized as follows:

- mismatch episodes are diagnosed via a taxonomy of within-model routes (parameter retuning, measurement/noise-model repair, precision reallocation) versus more radical change (between-model selection and constraint revision) [12, 26];
- hypothesis formation is model search and construction constrained by admissibility conditions, not merely parameter estimation within a fixed class [17, 30];
- discovery occurs when revision yields an inferentially and methodologically stabilized way of seeing-as and reporting, often reorganizing the space of admissible observations and explanations (C3) [16, 24];
- revolutions involve revisions of constraint packages and associated practices, schematically $(T, N, I) \mapsto (T', N', I')$, reshaping observation, explanation, and the interpretation of data [2, 6, 24].

This Bayesian/PP vocabulary is explanatory and bridge-oriented: it articulates role constraints on how mismatch can drive revision, while leaving normative questions about justification, objectivity, and progress to distinct philosophical assessment [18, 40].

6.4 Limitations and Future Directions

Although this study achieved a coherent bridge proposal, important limitations remain.

(1) Limits of the computational idealization

PP/FEP should be read as a family of modeling strategies rather than a settled neurobiological theory. The extent to which biological brains implement Bayesian inference or free-energy minimization remains

an open empirical question [4, 11, 20]. The present bridge therefore relied only on weak commitments: approximate inference under generative models, mismatch signals with revision-driving roles, and precision allocation.

(2) Open gaps: language, norms, and social practice

While we modeled observation reports as report policies, a fuller account would require deeper integration with pragmatics and philosophy of language. Moreover, communal norms and instrumentation were treated as explicit constraints (N, I) on reporting and stabilization rather than being reduced to neural computation; this is appropriate for a non-reductive bridge, but it leaves open the detailed mechanisms of institutional stabilization and collective epistemic dynamics [2, 6].

(3) AI as a testbed and a target

Modern AI implements approximate inference, model comparison heuristics, and representation learning, but robust algorithms for open-ended model-space revision (the analogue of constraint revision $T \mapsto T'$) remain limited relative to human scientific practice [1, 14, 28, 37]. The present framework also clarifies a target for AI/ML as a testbed: explicit model-space constraints, diagnostics for structured residuals, and policy-governed reporting under norms provide a concrete analogue of constraint revision in scientific practice [26, 37].

(4) Prospects for interdisciplinary integration

Despite these limitations, the bridge clarifies concrete points of contact:

- **Neuroscience:** precision mechanisms, hierarchical inference, and the conditions under which mismatch becomes revision-driving [11, 25, 31, 35];
- **AI/ML:** model search and meta-learning for representation and hypothesis revision [14, 26, 37];
- **Philosophy of science:** reinterpreting historical case studies via explicit role constraints and failure modes, and reassessing objectivity via data/phenomena extraction and public reporting norms [2, 6, 16, 24];
- **Cognitive science and psychopathology:** using precision-weighting failures as a principled locus for atypical perception, linking PP interpretations to broader issues of theory-ladenness [15, 23, 34, 35].

Concluding Remarks

Observation is not mere seeing, but inference arising from the interaction between generative models and sensory or instrumental data. In this process, theories—modeled as constraints on admissible model space—shape the form of seeing-as, and observation reports emerge as goal- and norm-sensitive report policies grounded in posterior states.

Discovery, in an important sense, is the structured management of misfit: persistent, precision-weighted residual patterns motivate within-model repairs, between-model selection, and, when required, revisions

of the admissibility constraints themselves. Science, in this view, is a long-term communal enterprise of model search and stabilization under shared measurement regimes and reporting norms, not a pipeline from "pure data" to theory.

Four anticipated objections (short replies).

(A) "PP/FEP is speculative; it cannot ground philosophy." The proposal does not treat PP/FEP as established truth. It relies only on weak commitments (approximate inference under generative models, mismatch as a revision-driving role, and precision allocation). When these fail, the bridge fails—hence explicit success conditions and failure modes (Section 1.4).

(B) "Theory-ladenness is semantic/normative; posteriors cannot capture it." We do not identify semantics with posteriors. Semantic and normative constraints enter via theory as admissibility constraints on model space and via report policies constrained by communal norms and measurement regimes (N, I), while the personal-level description retains its autonomy [2, 18, 32].

(C) "This is judgment, not perception." The target is scientific observation as a composite practice: hierarchical inference, precision allocation, and policy-governed reporting. Strong cognitive penetration of early vision is not required [9].

(D) "Circle is a special case of ellipse; so it is not between-model." The relevant contrast is a change in admissibility constraints (an eligibility regime that effectively enforces $e = 0$ versus one that treats $e \neq 0$ as admissible), i.e., a change in which models are eligible explanations, not merely a change in an already-admissible parameter value (Sections 1.5, 5.2) [12, 26].

A A Minimal Model-Comparison Sketch: Circle vs Ellipse

This appendix provides a deliberately minimal sketch of Bayesian model comparison for the Kepler case discussed in Section 5. Its purpose is not to reproduce Kepler's historical calculations, but to make explicit—in a toy setting—how (i) persistent *structured* residual patterns can exceed the capacity of within-model parameter adjustment and (ii) between-model selection (and, in the spike–slab reading below, admissibility-constraint revision) can be rationally motivated despite higher model complexity. The sketch supports the bridge criteria stated in Section 1.4, especially the distinction between within-model updating and between-model/constraint revision (cf. Section 5.2).

Scope of the toy sketch. The appendix is designed to make one structural point explicit: why persistent *structured* misfit can rationally favor relaxing an eligibility constraint (here, allowing $e \neq 0$) despite an Occam-style complexity cost. It is *not* intended as a realistic reconstruction of Keplerian dynamics, historical data processing, or the details of orbital time-to-angle mappings; those details are bracketed because they are not needed for the model-evidence/constraint-revision moral.

Data (idealization). Let $s = \{(t_i, y_i)\}_{i=1}^{n}$ denote a time-indexed observational record, where each y_i summarizes a measurement of a planet's apparent position at time t_i. For concreteness, take $y_i \in \mathbb{R}^2$ (e.g., $y_i = (x_i^{(1)}, x_i^{(2)})$ in a fixed coordinate system), though an angular encoding would work as well. We treat

measurement noise as approximately Gaussian with known scale σ. (Allowing σ to be unknown changes details but not the Occam-trade-off moral.)

Generative-model perspective. For each model class M with parameters θ, assume a likelihood $p(s \mid \theta, M)$ and a parameter prior $p(\theta \mid M)$. Model comparison proceeds via model evidence (marginal likelihood),

$$p(s \mid M) = \int p(s \mid \theta, M)\, p(\theta \mid M)\, d\theta,$$

and posterior model probability,

$$p(M \mid s) \propto p(s \mid M)\, p(M),$$

where $p(M)$ is a prior over model classes [12, 21, 26].

A note on "anomaly" in the toy setting. In line with discriminator D2 (Section 1.12), we treat an "anomaly" not as a single large deviation but as *persistent structured* residual patterns that survive best within-model fitting and plausible measurement/noise-model variants (cf. posterior predictive checking and residual diagnostics) [12].

A.1 Two Candidate Model Classes

We compare two idealized model classes.

Model class 1: circular orbit (M_{circle}). Let θ_c denote parameters (e.g., radius, phase, orientation, and a crude time-to-angle mapping). A simple geometric parametrization is:

$$\hat{y}(t; \theta_c) = R_\alpha \begin{pmatrix} r\cos(\omega t + \phi) \\ r\sin(\omega t + \phi) \end{pmatrix} + \tau_c,$$

where $r > 0$ is the radius, ω an angular rate, ϕ a phase, R_α a planar rotation by angle α, and $\tau_c \in \mathbb{R}^2$ a translation capturing coordinate centering. (Any comparable low-parameter "circle + adjustments" family would serve.)

Assume isotropic Gaussian measurement noise:

$$p(s \mid \theta_c, M_{\text{circle}}) = \prod_{i=1}^{n} \mathcal{N}\big(y_i;\, \hat{y}(t_i; \theta_c),\, \sigma^2 I\big)\,.$$

Define residuals as

$$\varepsilon_i(\theta_c) = y_i - \hat{y}(t_i; \theta_c).$$

Model class 2: elliptical orbit (M_{ellipse}). Let θ_e denote parameters that allow an ellipse:

$$\hat{y}(t; \theta_e) = R_\alpha \begin{pmatrix} a\cos u(t) + c_e \\ b_e \sin u(t) \end{pmatrix} + \tau_e, \qquad b_e = a\sqrt{1 - e^2}, \quad c_e = -ae,$$

where $a > 0$ is the semi-major axis, $e \in [0, 1)$ eccentricity, b_e the semi-minor axis, c_e an offset that breaks circular symmetry, $\tau_e \in \mathbb{R}^2$ a translation, and $u(t)$ is a toy phase parameterization. For minimality one may take $u(t) = \omega t + \phi$; in real Keplerian dynamics the time-to-angle mapping is not linear (Kepler's equation), but that dynamical detail is not needed for the model-evidence/Occam point here. Note that this parametrization recovers the circular case as the special case $e = 0$ (hence $b_e = a$ and $c_e = 0$), which supports the nested-model reading used below.

Assume Gaussian measurement noise:

$$p(s \mid \theta_e, M_{\text{ellipse}}) = \prod_{i=1}^{n} \mathcal{N}\big(y_i;\, \hat{y}(t_i; \theta_e),\, \sigma^2 I\big)\,,$$

with residuals $\varepsilon_i(\theta_e) = y_i - \hat{y}(t_i; \theta_e)$.

Within-model updating vs. structural mismatch (toy analogue of D2). Within a fixed model class M, within-model updating corresponds to selecting (or learning) parameters θ that improve fit, e.g. by maximizing $p(s \mid \theta, M)$ (maximum likelihood) or $p(\theta \mid s, M)$ (MAP). If, however, residuals retain *systematic structure* even at the best-fitting parameters—i.e., $\varepsilon_i(\hat{\theta}_c)$ remain patterned rather than noise-like across i—then the discrepancy is not merely parametric but indicates a mismatch in model class. This is the toy analogue of a persistent anomaly.

Where measurement/noise-model repair would enter (link to within-model(B)). In the main text (Section 5.2), many apparent anomalies dissolve once the measurement/noise model is repaired. In the present toy sketch this would correspond to revising the likelihood family (e.g., allowing heteroscedasticity, correlations, or non-Gaussian tails) or revising calibration assumptions. Only residual structure that survives such plausible likelihood variants counts as strong evidence for model-class (or constraint) revision [2, 12].

Precision-weighted error (schematic PP alignment). To align with the PP vocabulary (Section 3), one may treat the effective discrepancy driving revision as precision-weighted:

$$\tilde{\varepsilon}_i(\theta) = \Pi_i\, \varepsilon_i(\theta),$$

where Π_i encodes reliability/precision (possibly time- and level-dependent). Here Π_i is a schematic scalar weight for simplicity; in full PP formulations, precision control is typically level- and channel-dependent rather than a single per-datum factor. The key point is that high-precision data make systematic residual structure harder to dismiss as noise.

A.2 Model Evidence and the Occam Trade-off

Model selection compares M_{circle} and M_{ellipse} via:

$$\frac{p(M_{\text{ellipse}} \mid s)}{p(M_{\text{circle}} \mid s)} = \frac{p(s \mid M_{\text{ellipse}})}{p(s \mid M_{\text{circle}})} \cdot \frac{p(M_{\text{ellipse}})}{p(M_{\text{circle}})}.$$

Even if one begins with a strong prior preference for circles (i.e., $p(M_{\text{circle}}) \gg p(M_{\text{ellipse}})$), sufficiently larger evidence $p(s \mid M_{\text{ellipse}})$ can overturn that preference.

Why evidence is not just best fit. Evidence integrates fit over parameter space:

$$p(s \mid M) = \int p(s \mid \theta, M)\, p(\theta \mid M)\, d\theta,$$

which implements an Occam-style trade-off: a more flexible model can fit better, but it pays a complexity cost because only a smaller fraction of its parameter space yields that fit [12, 26].

Laplace/BIC-style approximation (optional heuristic). Around a posterior mode $\hat{\theta}$, one may use a Laplace approximation:

$$\log p(s \mid M) \approx \log p(s \mid \hat{\theta}, M) + \log p(\hat{\theta} \mid M) + \frac{k}{2} \log(2\pi) - \frac{1}{2} \log |H|,$$

where k is the number of parameters and H is the Hessian of $-\log p(\theta \mid s, M)$ at $\hat{\theta}$. Heuristically, the last term is an "Occam factor" penalizing overly flexible models unless improved fit is robust [26]. A cruder but widely used proxy is the BIC penalty, which highlights the same idea: improved fit must compensate for added complexity. This approximation is included only as intuition for the Occam factor; the bridge argument does not rely on BIC/Laplace per se, but on the general fact that evidence integrates fit over parameter space.

Nested-model reading (spike vs. slab on eccentricity; link to admissibility constraints). Because a circle is a limiting case of an ellipse ($e = 0$), one can express the contrast as a nested comparison by treating eccentricity explicitly. A "circle-only" admissibility regime can be idealized as a spike prior $p(e \mid M_{\text{circle}}) = \delta(e)$, whereas an "ellipse-admissible" regime uses a continuous (slab) prior $p(e \mid M_{\text{ellipse}})$ on $[0, 1)$. Here ϑ collects all remaining (non-eccentricity) parameters of the orbit/measurement model. The Bayes factor then compares:

$$\text{BF}_{\text{ellipse,circle}} = \frac{p(s \mid M_{\text{ellipse}})}{p(s \mid M_{\text{circle}})} = \frac{\int p(s \mid e, \vartheta, M_{\text{ellipse}})\, p(e, \vartheta \mid M_{\text{ellipse}})\, de\, d\vartheta}{\int p(s \mid \vartheta, M_{\text{circle}})\, p(\vartheta \mid M_{\text{circle}})\, d\vartheta}.$$

On this reading, the "between-model" shift corresponds to revising the admissibility constraint from a spike at $e = 0$ to a slab that allows $e \neq 0$. This is a toy formalization of the constraint-revision point made in Section 1.5: the change is not merely parametric but alters what counts as an eligible explanation.

Interpretive moral for the Kepler case (toy form). First, one performs within-model fitting in M_{circle}. If residuals remain systematically structured rather than noise-like at best fit (and under plausible measurement/noise-model variants), then the discrepancy is not merely parametric but indicates a mismatch in model class. Second, one compares evidence for an alternative class M_{ellipse}. If M_{ellipse} removes the structured residuals so that residuals are closer to noise (especially under high precision), then its evidence can dominate even after the Occam penalty, motivating a transition to the new admissibility regime.

Bridge connection. This appendix illustrates the bridge logic in minimal form: (i) an anomaly corresponds to persistent, precision-weighted structured residual patterns that resist within-model updating (and plausible measurement repair); (ii) discovery corresponds to adopting a new admissible model class (or revising admissibility constraints) whose evidence outweighs the old class despite complexity penalties. This is the minimal computational sense in which the observation–anomaly–discovery pattern can be rendered as multi-tier revision dynamics consistent with the bridge criteria (Section 1.4).

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[2] J. Bogen and J. Woodward, "Saving the phenomena," *Philosophical Review*, 97(3):303–352, 1988.

[3] A. Clark, "Whatever next? Predictive brains, situated agents, and the future of cognitive science," *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[4] A. Clark, *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, Oxford University Press, 2015.

[5] C. F. Craver, *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*, Oxford University Press, 2007.

[6] P. Dear, *The Intelligibility of Nature: How Science Makes Sense of the World*, University of Chicago Press, 2006.

[7] D. C. Dennett, *The Intentional Stance*, MIT Press, 1987.

[8] P. Feyerabend, *Against Method*, Verso, 1975.

[9] C. Firestone and B. Scholl, "Cognition does not affect perception: Evaluating the evidence for 'top-down' effects," *Behavioral and Brain Sciences*, 39:e229, 2016.

[10] K. Friston, "A theory of cortical responses," *Philosophical Transactions of the Royal Society B*, 360(1456):815–836, 2005.

[11] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, 11:127–138, 2010.

[12] A. Gelman et al., *Bayesian Data Analysis*, 3rd ed., CRC Press, 2013.

[13] D. Gillies, *Philosophical Theories of Probability*, Routledge, 2000.

[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[15] R. L. Gregory, "Perceptions as hypotheses," *Philosophical Transactions of the Royal Society B*, 290(1038):181–197, 1980.

[16] N. R. Hanson, *Patterns of Discovery*, Cambridge University Press, 1958.

[17] G. Harman, "The inference to the best explanation," *Philosophical Review*, 74(1):88–95, 1965.

[18] C. G. Hempel, *Aspects of Scientific Explanation*, Free Press, 1965.

[19] W. E. Hill, "My wife and my mother-in-law. They are both in this picture–find them," *Puck*, week ending Nov. 6, 1915, p. 11.

[20] J. Hohwy, *The Predictive Mind*, Oxford University Press, 2013.

[21] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, 2003.

[22] D. M. Kaplan and C. F. Craver, "The explanatory force of dynamical and mathematical models in neuroscience: a mechanistic perspective," *Philosophy of Science*, 78(4):601–627, 2011.

[23] D. Kersten and A. Yuille, "Bayesian models of object perception," *Current Opinion in Neurobiology*, 13(2):150–158, 2003.

[24] T. S. Kuhn, *The Structure of Scientific Revolutions*, University of Chicago Press, 1962.

[25] T. Lee and D. Mumford, "Hierarchical Bayesian inference in the visual cortex," *Journal of the Optical Society of America A*, 20(7):1434–1448, 2003.

[26] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[27] D. M. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W. H. Freeman, 1982.

[28] K. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012.

[29] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000.

[30] C. S. Peirce, *Collected Papers of Charles Sanders Peirce*, Vols. 1–6, Harvard University Press, 1931–1935.

[31] R. Rao and D. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nature Neuroscience*, 2(1):79–87, 1999.

[32] W. Sellars, "Empiricism and the philosophy of mind," *Minnesota Studies in the Philosophy of Science*, 1:253–329, 1956.

[33] A. K. Seth, "A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia," *Cognitive Neuroscience*, 5(2):97–118, 2014.

[34] A. K. Seth, *Being You: A New Science of Consciousness*, Faber & Faber, 2021.

[35] P. Sterzer et al., "The predictive coding account of psychosis," *Biological Psychiatry*, 84(9):634–643, 2018.

[36] D. Stokes, "Cognitive penetration and the perception of art," *Dialectica*, 68(1):1–34, 2014.

[37] J. B. Tenenbaum et al., "How to grow a mind: Statistics, structure, and abstraction," *Science*, 331(6022):1279–1285, 2011.

[38] J. R. Voelkel, *The Composition of Kepler's Astronomia Nova*, Princeton University Press, 2001.

[39] R. S. Westfall, *Force in Newton's Physics*, Macdonald and Co.; American Elsevier, 1971.

[40] J. Woodward, *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press, 2003.