The Epistemology of Observation: A Formal Certification Framework for Human and Automated Classifiers in DAG-OR

Vincent Gonzalez

Independent Scholar · ORCID 0009-0004-7283-8598

modulign.org · VinceGonzalez@me.com

Abstract

Every classification system presupposes an observer. The Modulign Standard (DAG-OR) makes this presupposition explicit through the §OBS and §AUT segments of the dimensional address, which encode not merely who observed but at what level of certified competence. This paper provides the formal academic treatment of the %AUT Certification Framework, previously published as an appendix to the Modulign v3.1 specification [7]. We develop the epistemological foundations for a three-tier observer competence model—:T (Trained), :E (Evaluated), :A (Autonomous)—and argue that observer certification is not an administrative overlay on the classification grammar but a constitutive requirement of it. We specify the enforcement mechanisms for prohibited practices, the minimum evaluation requirements for each certification tier, the explainability requirements and their limits, the out-of-distribution detection mandates, and the constitutional constraints on automated observer testimony. We conclude with an honest assessment of the framework's limitations, including the certifier competence recursion and the conditions under which the framework would require revision.

Keywords: observer certification, epistemic competence, automated classification, explainability, AI governance, Confrontation Clause, dimensional address grammar

1. Introduction: The Observer Problem

Classification is not a view from nowhere. Every act of classification—naming a specimen, tagging an evidence sample, encoding a measurement—is performed by someone or something, under specific conditions, with specific competencies and specific limitations. The history of science is, in significant part, the history of learning to take the observer seriously: the astronomer's personal equation, the experimenter's bias, the instrument's calibration curve, the algorithm's training distribution.

Yet most classification systems treat the observer as external metadata. The specimen label records what was classified, perhaps when and where, but the classifier's competence, the instrument's calibration status, and the conditions under which the observation was made are documented—if at all—in separate records that may or may not travel with the classification. The chain from observation to observer to competence to limitation is, in practice, routinely broken.

The Modulign Standard [1] addresses this structurally. The §OBS segment encodes the observer's identity. The §AUT segment encodes, for non-human observers, the observer's certification level

under the %AUT framework [7]. Together, these segments make the observer a first-class component of the dimensional address—not an annotation on the observation but a constitutive dimension of it. This paper provides the full academic treatment of that framework: the arguments for why observer competence must be encoded this way, what the three certification tiers mean epistemically, and where the framework honestly falls short.

2. The Argument for Constitutive Observer Encoding

2.1 The Constitutivity Claim

The central claim of this paper is that observer competence is not a property of the observer that happens to be relevant to classification. It is a constitutive dimension of the classification itself. A measurement made by a calibrated instrument and the same measurement made by an uncalibrated instrument are not the same measurement with different metadata—they are different measurements. A classification produced by a trained human expert and the same classification produced by an untrained layperson are not the same classification with different confidence levels—they are classifications with different epistemic standing, and the address must reflect this.

This claim follows from the ATNA principle proven in the Formal Logic paper [2]: Addressability Tracks Noetic Access. If the dimensional address encodes the full epistemic access conditions under which an observation is known, and if the observer's competence is a determinant of epistemic access, then observer competence must be encoded in the address. Omitting it would violate ATNA—the address would claim to specify access conditions that it does not in fact specify.

2.2 Dual-Mechanism Constitutivity

The constitutivity of observer competence operates through two distinct mechanisms depending on whether the observer is human or automated. For automated observers, competence is encoded explicitly in §AUT—the three-tier certification (:T, :E, :A) directly populates the address with the observer's competence level. For human observers, competence is encoded implicitly through the CDP itself: a human classifier operating under the eight-step protocol is constrained by the protocol's decision tree, boundary rules, and mandatory segment requirements. Deviations from the CDP are detectable in the classification output. The protocol is the competence framework for human observers.

The §AUT segment is therefore conditional—populated only for non-human observers—without violating the constitutivity claim. For humans, the CDP does the work. For machines, §AUT does the work. This asymmetry has a specific justification: human observers can be cross-examined about their reasoning; automated systems cannot. The CDP constrains what the observer does. The %AUT certification constrains what the observer is competent to do. For human observers, cross-examination can test competence independently of the address. For automated systems, the address must encode competence because no other mechanism is available to surface it.

2.3 Why Three Tiers

:T (Trained)—The system has been trained on the CDP and can execute it within a restricted domain. A :T system can follow the protocol. It cannot evaluate whether the protocol applies. :E (Evaluated)—The system has been evaluated against Inter-Rater Reliability benchmarks [13] and

its error characteristics are known and documented. The distinction from :T is not merely that the system has been tested—it is that its failure modes are characterized. :A (Autonomous)—The system is certified for unsupervised classification with mandatory out-of-distribution detection and explainability requirements. An :A system must know when it doesn't know. OOD detection is not optional at this level because autonomous operation means no human is available to catch inputs that fall outside the training distribution. An :A system that cannot detect distribution shift is, by the framework's definition, not autonomous—it is merely unsupervised, which is a different and more dangerous thing.

The three tiers are ordered by epistemic self-awareness: the degree to which the system can characterize its own competence boundaries. :T systems classify. :E systems classify with known error. :A systems classify with known error and can detect when their error model no longer applies.

3. Minimum Evaluation Requirements

3.1 :T Certification Requirements

A :T certification requires: demonstrated ability to execute the CDP on inputs within a specified domain restriction; a defined domain scope identifying which domains, realms, and scale tiers the system is certified to classify; and a training data provenance record. A :T certification does not require a measured error rate. This is the defining limitation of the :T tier: the system may work correctly, but its reliability is not quantified.

3.2 :E Certification Requirements

A :E certification requires everything a :T certification requires, plus: a minimum test-set size of at least 1,000 classified observations or 10× the number of domain/realm/scale combinations in scope, whichever is larger; mandatory stratification across all combinations with a minimum of 30 observations per stratum; published error characterization including false-positive rates, false-negative rates, and confusion matrices by domain/realm; and a 12-month recertification period or upon any specification amendment changing domain boundaries, whichever is sooner. Publication of results is a condition of certification, not a recommendation.

3.3 :A Certification Requirements

A :A certification requires everything a :E certification requires, plus: OOD detection evaluation against defined distributional shifts (in-distribution, near-distribution, and out-of-distribution inputs); a measured and published OOD false-negative rate below the domain-specific threshold (proposed: 5% forensic, 10% scientific, 15% general); and an explainability evaluation in which protocol traceability records are reviewed by a human expert who confirms the traces are meaningful rather than superficially plausible. Every :A certification carries a scope limitation specifying the classes of distributional shift against which detection has been validated.

4. Explainability Requirements

4.1 What Explainability Means in This Context

Within the framework, explainability has a precise definition: for any classification produced by a :E or :A system, the system must be able to produce a trace mapping the classification to specific CDP steps—identifying which inputs triggered which domain, realm, and scale decisions, and which segment values were populated on the basis of which features. This is protocol traceability, not general model interpretability. The narrowness is deliberate: protocol traceability is a well-defined engineering requirement. General interpretability of a neural network's internal representations is an open research problem without consensus solutions.

4.2 The Explainability Ceiling

Protocol traceability does not explain why the system made a particular domain assignment at CDP Step 2. It explains that the system made it, and what input features were associated with it. For a black-box classifier, the mapping from input features to CDP decision may itself be opaque. This ceiling must be disclosed. A cross-examining attorney will ask: “You can tell me what the system did. Can you tell me why it did it?” The honest answer, for many :E-level systems, is that the internal computation cannot be fully explained. The Human Attestation Protocol [20] specifies how this ceiling is handled in testimony.

4.3 The Audit Function

Protocol traceability serves a specific function in the Modulign architecture: it enables audit. If a classification is contested, the traceability record provides the basis for evaluating whether the CDP was correctly applied. Without traceability, the only recourse is to reclassify from scratch. With traceability, the specific step at which disagreement arose can be identified, and the Correction Protocol [19] can reference the precise basis for correction.

5. Out-of-Distribution Detection

5.1 The Epistemic Obligation

The requirement for OOD detection at the :A level is the framework's most demanding technical specification. The ATNA principle requires that the dimensional address encode the epistemic access conditions for the observation. A classification produced outside the system's evaluated distribution has unknown epistemic standing—the error model no longer applies, the reliability benchmarks no longer hold, the §CONF value is meaningless. Issuing a classification under these conditions would populate the address with values that misrepresent the epistemic situation. OOD detection is therefore not a safety feature; it is an epistemic obligation.

5.2 Implementation Constraints

OOD detection is an active research area without settled solutions. The framework specifies a performance requirement, not a method: the :A-level system must demonstrate OOD detection capability against defined distributional shifts with a false-negative rate below the domain-specific threshold. Every :A certification must specify the classes of distributional shift against which detection has been validated, because detection cannot be claimed universally. Novel input types

may fall outside the OOD detector's own training distribution. The scope limitation is the honest acknowledgment of this.

6. Prohibited Practices and Enforcement

6.1 Prohibited Practices

Four practices are formally prohibited for automated classifiers operating under DAG-OR: (1) Classification without provenance—violates Invariant 9; (2) Classification with suppressed confidence intervals—violates ATNA; (3) Classification of persons without consent—violates the Privacy Commitment [8]; (4) Retroactive competence claims—violates Invariant 2 (Permanence). The §AUT value at commit is permanent.

6.2 Enforcement Mechanisms

Classification without provenance is enforced by rejection at commit: an incomplete §PROV chain fails the mandatory-segment check and the classification is not committed. This is a database constraint, not a policy. Classification with suppressed confidence intervals is detected by periodic audit: a §CONF value populated but undocumented, or a value of exactly 1.0 from a system with known error rates, is flagged as presumptively erroneous and queued for review under the Correction Protocol [19]. Classification of persons without consent is held for human review before commit when input metadata indicates identifiable persons. Retroactive competence claims are rejected by the append-only constraint itself—the architecture enforces this prohibition, not a policy document.

7. The Certifier Problem

7.1 Who Certifies the Certifier

The framework specifies that automated observers must be certified. The immediate question is: by whom, with what authority, and how is the certifier's own competence established? The current specification (Phase 1 governance [8]) designates the sole author as the certification authority. This is adequate for a research-stage system. It is not adequate at scale. The governance transition [8, 18] addresses this: Phase 2 introduces an advisory board with domain experts capable of evaluating certifications; Phase 3 establishes distributed certification authority subject to cross-recognition protocols. The honest statement is that the certification framework is currently as strong as its sole certifier, and no stronger.

7.2 Certifier Liability

When the certifier certifies a system that subsequently produces harmful errors, potential liability arises—negligent certification, negligent misrepresentation, product liability. The certification instrument should specify that certification attests to evaluation results at the time of certification, not to future performance. The Phase 2 governance entity should carry professional liability insurance for certification activities. Until these mitigations are in place, the sole author certifies at personal risk. This is stated because it is real and should be visible.

7.3 Comparison with Existing Frameworks

ISO/IEC 17025 specifies competence requirements for testing and calibration laboratories—more developed institutionally, less specific epistemically. A mature %AUT implementation would benefit from ISO 17025's institutional model. The EU AI Act conformity assessment classifies AI systems by risk level; %AUT classifies observers by epistemic competence within DAG-OR. The two frameworks are complementary, not redundant—an automated Modulign classifier deployed in the EU must satisfy both independently. NIST AI RMF 1.0 is a voluntary risk management framework; %AUT mandates. The two are compatible and address different scopes.

8. Inter-Rater Reliability and Human Certification

The Inter-Rater Reliability Study [13] measured 44.4% overall agreement across all domains, with domain-specific rates substantially higher (TRN = 100%). The v3.1 amendment [14] was motivated by the NAT/ENV boundary disagreement identified in that study. The relevant question for this paper is: at what point does inter-rater reliability data indicate that the CDP alone is insufficient as a competence framework for human observers?

The threshold should be specified: if, after specification amendments addressing identified boundary ambiguities, the overall inter-rater agreement rate for human classifiers trained on the current CDP version remains below 70% across two successive studies with $N \geq 200$, the framework should introduce formal human certification tiers analogous to the automated tiers. The 70% threshold reflects the conventional benchmark for substantial agreement (Cohen's kappa $\geq$ 0.61). The CDP's sufficiency is an empirical hypothesis, not an axiom.

Good-faith disagreement between competent classifiers on boundary cases is data—it informs specification refinement and reveals where domain boundaries need sharpening. Classification error, by contrast, occurs when the CDP is misapplied. Both generate corrections under the Correction Protocol [19], but they are tracked differently: boundary disagreements are recorded with relation type DIVERGES, contributing to the inter-rater reliability dataset that drives specification refinement; classification errors are recorded as Type 1 or Type 2 corrections with a stated basis identifying the CDP step at which misapplication occurred. The distinction is not always clean—some corrections will be reclassified as the specification evolves—but the Correction Protocol accommodates this through its dispute mechanism.

9. Open Problems and Falsification Conditions

The framework would require fundamental revision if: (1) The three-tier model proves insufficiently granular for a significant class of automated observers. (2) OOD detection proves fundamentally intractable for classification-relevant distributional shifts—meaning autonomous unsupervised classification is not currently achievable, an honest conclusion if the evidence demands it. (3) The certifier recursion problem is not resolved by the governance transition. (4) The CDP proves insufficient as a competence framework for human observers, requiring human certification tiers. (5) The enforcement mechanisms specified in §6.2 prove inadequate for emerging violation classes—real-time statistical monitoring of classification distributions may be required.

10. Conclusion

The observer is not optional. Every classification presupposes one, and the epistemic standing of the classification depends on the competence of the observer that produced it. The %AUT Certification Framework makes this dependence explicit, formal, and auditable. It defines what it means for an automated system to be competent to classify within DAG-OR, at three levels of epistemic self-awareness, with specific requirements for evaluation, traceability, error characterization, distributional shift detection, and enforcement of prohibited practices.

The framework is young. Its institutional infrastructure is minimal. Its certifier authority is centralized in a single individual. These are limitations, stated honestly, with a governance roadmap that addresses them. What exists today is the formal specification of what observer certification means within a dimensional address grammar—and the argument that any classification system that does not encode observer competence as a constitutive dimension of the classification is missing something that matters.

References

[1] Gonzalez, V. (2026). *The Modulign Standard v3.0*. Zenodo. https://doi.org/10.5281/zenodo.19348703

[2] Gonzalez, V. (2026). *Modulign Formal Logic v1.1.0*. Zenodo. https://doi.org/10.5281/zenodo.19350847

[7] Gonzalez, V. (2026). *%AUT Certification Framework (Modulign v3.1 Appendix)*. Zenodo. https://doi.org/10.5281/zenodo.19559602

[8] Gonzalez, V. (2026). *Privacy Commitment & Governance Roadmap v1.0.0*. Zenodo. https://doi.org/10.5281/zenodo.19559689

[13] Gonzalez, V. (2026). *Inter-Rater Reliability in Modulign Classification: An Initial Study*. Zenodo. https://doi.org/10.5281/zenodo.19643322

[14] Gonzalez, V. (2026). *Modulign v3.1 Amendment: ENV and BIO Domains*. Zenodo. https://doi.org/10.5281/zenodo.19644043

[18] Gonzalez, V. (2026). *Federated Classification Infrastructure: Registry Governance, Node Architecture, and Decentralized Stewardship for DAG-OR*. Zenodo. https://doi.org/10.5281/zenodo.19726393

[19] Gonzalez, V. (2026). *The Modulign Correction Protocol: Formal Specification for Append-Only Error Resolution in DAG-OR*. Zenodo. https://doi.org/10.5281/zenodo.19726401

[20] Gonzalez, V. (2026). *The Human Attestation Protocol: Confrontation Clause Compliance for Automated Classifications in DAG-OR*. Zenodo. https://doi.org/10.5281/zenodo.19726407