The Universal Fallacy Architecture

Structural Error Prediction from AOML Constraint Geometry

Brent Alan Seeley · WellFlow Labs

https://orcid.org/0009-0002-9902-5955 · bseeley@wellflowlabs.com

Version: August 2026 Preprint (Paper II) · Series: Foundations of the Convergent Semantic Architecture (Paper II of IV)

DOI: https://doi.org/10.5281/zenodo.21998398

Abstract

This paper formalizes the Universal Fallacy Architecture (UFA), a predictive constraint system formalizing the semantic constraint geometry first mapped by Cross-Linguistic Compositional Analysis (CLCA, Paper I) and independently evidenced by the fallacy tradition itself: logical fallacies are not arbitrary errors but systematic violations of a finite set of semantic boundaries. CLCA reveals that human languages, despite surface diversity, converge on a shared semantic geometry governing how truth claims may be composed without collapse. These constraints are formalized in the Axis–Operator–Method–Level (AOML) framework.

We show that all stable structural reasoning errors arise from a single failure mechanism, Mode-Condition Collapse, in which analogical pattern transfer preserves relational structure while violating required semantic conditions. Exhaustive analysis demonstrates that all classical, modern, and rhetorical fallacies reduce to five fundamental boundary violations: Axis Transfer Error, Operator Overextension, Verification Habitat Mismatch, Stratification Failure, and Contrast Frame Collapse. The claim is exhaustiveness over structural reasoning failure (cases in which the distinctions being drawn are themselves malformed), not over quantitative failure such as miscalibrated credence or performance failure such as attention lapse (§2.1).

This revision builds on AOML v2.2, with REVEAL resolved from provisional to core operator status (see the operator discussion in §1.2), a three-layer constraint architecture that separates primitive geometry (six primary axes, four meta-axes, seven core operators, four levels) from empirically discovered structural rules and validation habitat constraints. Its central result is the meta-axis substitution family: a single schema, M[A] ↛ A, instantiated on each of the four meta-axes (DEG, TEMP, LIKE, EVID) by rules R5, R6, R8, and R9, generating a large and principled family of fallacies (Appeal to Probability, Appeal to Novelty, Forgery, Sophistry, Truthiness, Hypocrisy, Demagoguery, and the evidential appeals) as a single Type 2 phenomenon. Further additions include: polarity-conditioned operator licensing (the CAUSE(+)/CAUSE(−) distinction on CAUSE, R3, and the agentive polarity asymmetry on ABSTR, R7); and

target-type classes (FACTUAL, EPISTEMIC, DISPOSITIONAL) as a candidate deeper layer that may unify the three strongest CLCA findings under a single structural constraint: a layer independently anticipated by the causative-typology literature (Shibatani 1976; Dixon 2000; Wolff 2003).

The framework makes explicit, falsifiable predictions across reasoning, translation, and artificial intelligence domains. The resulting classification is not a taxonomy but a structural projection of the AOML constraint space: once the catalog of axes, operators, methods, and levels is specified, the space of possible structural errors is calculable from it, yielding a universal fallacy predictor.

1. The Analogy Paradox: Foundation of Structural Error

1.1 Analogy as the Engine of Insight and Error

Paper I established that unrelated languages converge on the same semantic architecture when you decompile their truth-related compositional patterns. This paper asks the next question: what happens when that architecture is violated?

Evidence structure. The claims of this paper rest on two distinguishable tiers of evidence. Tier 1 is the constraint geometry in reasoning systems: that the AOML geometry is real, finite, and load-bearing for reasoning is supported by three lines independent of how the geometry was first discovered: (i) the Universal Fallacy Map: 101 classical and contemporary fallacies, named by the human reasoning-error tradition over 2,400 years, decompose cleanly into the five violation types with no sixth type required; (ii) oAOML (operational AOML): engineering failure analyses exhibiting the same violation structure in a non-linguistic substrate (the engineering companion, https://doi.org/10.5281/zenodo.21960566); and (iii) structural-determinacy results: on a hand-adjudicated contested-boundary corpus, frontier language models given a claim's bare axis-tuple systematically fail to derive the map-canonical violation type, while the AOML structural signals deterministically recover it, evidence that the framework contributes discriminating structure a frontier model does not already possess.

Tier 2 is the geometry as a human-language universal: that the same geometry is convergently encoded in the morphology and composition of human languages is the claim of Paper I (CLCA). It is supported by cross-linguistic elicitation across sixteen languages and is undergoing native-speaker validation (Paper I, Appendix B); pending that validation, Paper I presents its regularities as candidates. The architecture of this paper requires Tier 1. Tier 2, if it survives validation, establishes something stronger: that the constraint geometry is not an artifact of any one reasoning substrate but a convergent property of distinction-making systems (the argument completed in Paper III). The two tiers fail independently and are evidenced independently.

Human reasoning achieves creative insight primarily through analogy, the ability to transfer structural patterns from one semantic domain to illuminate another. This is the engine of discovery: the planetary model of the atom, time as money, the reverse-compiler analogy itself. But this same mechanism is also the primary source of

systematic reasoning errors. The pattern transfers, but the conditions under which it is valid do not. That tension between insight and error is the subject of this paper.

1.2 The Failure Mechanism: Mode-Condition Collapse

Using the AOML primitives established in CLCA (Paper I) and formalized in the AOML v2.2 architecture, the structural danger of analogy can be precisely characterized. Analogy transfers Pattern-Correspondence (the structural relationship) but frequently fails to transfer the corresponding Manifestation-Mode or Verification-Method.

A fallacy is the result of this failure, a Mode-Condition Collapse, where the validation requirements of the source domain are improperly applied to the target domain.

The mechanism in coordinates. Consider the analogy at the root of social Darwinism. Source: competition drives adaptation in nature. Target: therefore competition ought to govern policy. The analogy is genuinely sound at the level of Pattern-Correspondence: the relational structure (pressure selects, selection improves) really does map from one domain to the other, and that correspondence is what makes the inference feel compelling. Written as tuples, the transfer becomes visible as a movement in coordinate space. The source claim is (VER, CAUSE, OBSERVATION, EMP): a veridical causal claim, validated by observation, at the empirical level. The target claim is (NORM, CAUSE, OBSERVATION, EMP). The axis has moved from VER to NORM (that movement is the analogy) while the operator, method, and level have been carried along unchanged. But the carried coordinates were licensed for the source position, not the destination. The claim's Manifestation-Mode has changed, a description has become a prescription, and the Verification-Method that grounded the source tuple grounds nothing at the target: normative claims are not validated by observation, and the empirical level grounds a claim that now lives at the conventional one. Pattern-Correspondence transferred; Manifestation-Mode and Verification-Method did not. That is Mode-Condition Collapse, located in the un-updated coordinates, and this instance has a classical name: the Naturalistic Fallacy (Type 1, Axis Transfer Error, §2.1). Every fallacy in this paper is such an event, an analogical movement in coordinate space that carries validation requirements the destination does not honor; the five types of §2 classify which condition failed.

AOML v2.2 Architecture Reference

AXIS (6 primary)	OPERATOR (8 core)	METHOD (8)	LEVEL (4)	META-AXIS (4)
VER (Veridical)	ATTR (Attribution)	OBSERVATION	EMPIRICAL	DEG (Degree): confirmed
AUTH (Authenticity)	CAUSE (Causation)	EXPERIMENT	CONVENTIONAL	TEMP (Temporal): confirmed
SINC (Sincerity)	BECOME (Becoming)	PROVENANCE	FORMAL	LIKE (Similative): confirmed
NORM	RESULT	TESTIMONY	ULTIMATE	EVID

AXIS (6 primary)	OPERATOR (8 core)	METHOD (8)	LEVEL (4)	META-AXIS (4)
(Normative)				(Evidentiality): confirmed
ESS (Essence)	ABSTR (Abstraction)	LOGIC		Target Types (Layer 2):
VERIF (Verification)	STRAT (Stratification)	STATISTICS		FACTUAL
	CONTR (Contrast)	CONSENSUS		EPISTEMIC
	REVEAL (core; formerly provisional)	PRINCIPLE		DISPOSITIONAL

Note: AOML v2.2 separates the architecture into three layers: primitive geometry (axes, operators, meta-axes, levels), structural rules (empirical constraints on the core legality matrix, including polarity conditioning and meta-axis substitution prohibitions), and validation habitats (axis→method mappings). This three-layer separation ensures that the primitive set remains minimal while the full explanatory power of the constraint system is preserved. The full specification is archived as the AOML v2.2 reference (https://doi.org/10.5281/zenodo.21792514).

v2.2 carries four meta-axes, all now cross-linguistically attested. DEG (degree) is attested across all sixteen languages, surfacing in CLCA-Revision and CLCA-1.1 alike. TEMP (temporal) was provisional in v2.1 under cross-protocol attestation alone; v2.2 gives it CLCA footing through reliability (a sustained-over-time modulation of sincerity (TEMP[SINC]) robustly lexicalized in both saturation sets) and it remains independently confirmed by the fallacy evidence (Appeal to Novelty, Chronological Snobbery, Sunk Cost, Historian's Fallacy, Presentism, and the temporal component of Appeal to Tradition all require TEMP as an active meta-axis). LIKE (the similative: "resembles A without being A") is the v2.2 reanalysis of what v2.1 listed as the LIFE axis: verisimilitude is LIKE[VER] and lifelikeness its aesthetic subset, but the same similative modulates every host axis (LIKE[AUTH] = forgery, LIKE[SINC] = hypocrisy, LIKE[VERIF] = sophistry, LIKE[NORM] = demagoguery, LIKE[VER] = truthiness, LIKE[ESS] = illusion), and it is grammaticalized via a truth-stem-plus-similative compound in six unrelated traditions (Latin vēri-similis, Greek αληθοφανής, Sanskrit -ābhāsa, Chinese 逼真, Tagalog parang totoo, Quechua cheqa hina). EVID (evidentiality) marks the source of a claim (direct, reported, or conjectural), grammaticalized three-way in Quechua (-m/-mi, -s/-si, -chá) and lexical across the sample.

REVEAL is the one operator whose evidentiary history we report in full, because its instruments required correction. Revelation constructions first surfaced organically in the naive arm's discovery phase, whose prompts contain no reveal-family vocabulary; later systematization steps in both pipelines, however, named the category, so attestation counts from those layers cannot certify emergence. The claim was therefore

retested under a pre-registered, deletion-only de-leaking of the guided pipeline: every reveal-family category name, host label, and exemplar verb was machine-deleted from the instruments, with the prompt diffs published in the replication package. With no reveal vocabulary available, dedicated revelation lexicons emerged in 16 of 16 languages for the original informant family and 14 of 16 for a second model family (Greek and Quechua returned none), with qualitative signatures that category-guided elicitation cannot produce: one informant filed Chinese 揭露 and 揭示 under an improvised OTHER classification because no offered category fit, and the second family volunteered the idiom 真相大白 (“the truth comes to light”) unprompted. The finding is consonant with the KNOWING IS SEEING mapping documented across languages (Sweetser 1990). The cross-architectural validation this operator's provisional status awaited is complete: REVEAL stands as a core operator of the v2.2 inventory. Two model families answer the single-model objection, not the shared-training-substrate objection, which the Appendix B native-speaker validation is designed to address.

The opposition topology (tokened the FRAME in the AOML v2.2 reference) is the precondition that makes the operator-licensing matrix meaningful, not a coordinate operator; Paper III formalizes this precondition status, and Type 5 violations (Contrast Frame Collapse) correspond to its mis-specification. CONTR, negative/contrastive attribution (negation), is a distinct coordinate operator, whose own misuse is a Type 2 violation; conflating the two under one label is the error this separation corrects.

1.3 The Refractive Solution

Effective reasoning is a constant negotiation between Analogical Projection, which generates insight by transferring structural patterns, and Refractive Boundary Detection, which preserves validity by enforcing semantic constraints. The term refractive denotes the process by which an incoming pattern is evaluated against a structured boundary, analogous to light bending as it passes into a medium of different density, to determine whether its manifestation mode, truth axis, and validation method remain coherent in the target domain.

In this framework, the AOML constraints function as refractive laws: invariant conditions that regulate how patterns may pass between semantic domains without distortion. When these constraints are respected, analogy yields insight; when they are violated, Mode-Condition Collapse occurs, producing systematic reasoning error.

A clarification on the metaphor: “refractive” is an analogy for the constraint-checking process, not a formal model of it. The formal mechanism is Mode-Condition Collapse, defined precisely in §2.2 in terms of the AOML tuple. The refractive metaphor is useful because it captures the intuition that patterns change character when they cross semantic boundaries, just as light changes behavior when it enters a new medium. But the mathematical properties of optical refraction (Snell’s law, refractive index) do not map one-to-one to the structural properties of AOML. The constraint geometry stands on its own formal definition; the metaphor is illustrative, not constitutive.

Example: Constraint Enforcement on Claims

Claim	AOML Tagging	Outcome	Constraint Violated
“Water boils at 100°C”	[VER.RESULT.EXPERIMENT.EMP]	✓ Valid	None
“Democracy is best”	[NORM.ATTR.CONSENSUS.CONV]	Requires normative reasoning	None (but requires normative process)
“Document is real”	[AUTH.RESULT.PROVENANCE.EMP]	✓ Valid	None
“Make someone sincere”	[SINC.CAUSE.*.*]	✗ Blocked	R1: Dispositional Causation Block
“Very likely, therefore true”	DEG[VER] → VER	✗ Blocked	R5: Meta-Axis Substitution

2. From Cross-Linguistic Constraints to a Universal Fallacy Architecture

The Universal Fallacy Architecture is not a post hoc classification of historical logical errors. Its five-type structure is fixed by the AOML constraint dimensions; that the 101-fallacy inventory decomposes into exactly these types, with every rule refinement landing inside an existing type, is the architecture’s primary evidence, and the CLCA discovery history (Paper I) is how the dimensions were first identified, not the sole ground on which they rest. If CLCA is the reverse compiler that recovers the instruction set, UFA is the error manual: a catalogue of what goes wrong when programs violate the instruction set’s rules. The key insight is that the catalogue is finite and predictable, because the instruction set is finite and structured.

CLCA demonstrates that human languages, despite radical surface diversity, converge on a shared semantic geometry governing how meaning may be composed without collapse. Logical fallacies arise when analogical reasoning transfers pattern correspondence across domains while violating one or more of those constraints. Fallacies are not arbitrary mistakes in reasoning style; they are predictable failure modes that occur when meaning is composed in ways that violate the stable geometry required for semantic coherence.

The discovery path ran as follows: CLCA identified candidate semantic axes (VER, AUTH, SINC, NORM, ESS, VERIF), each corresponding to a distinct kind of truth claim with non-interchangeable validation requirements. Compositional operators (ATTR, CAUSE, BECOME, RESULT, ABSTR, STRAT, CONTR, and REVEAL) are shown to be selectively licensed across axes, with certain combinations universally blocked. Validation methods define habitats appropriate to each axis. Stratification levels

(EMPIRICAL, CONVENTIONAL, FORMAL, ULTIMATE) impose constraints on permissible inference between instances and categories. Opposition topology, the precondition under which axis distinctions can be drawn at all, is the layer that the operator-licensing matrix presupposes; its mis-specification produces Type 5 violations and is formalized in Paper III as the precondition that makes the four AOML conditions possible. Together, these constraints define a finite semantic legality space. Violations of this space yield a corresponding finite set of failure modes. The axes' standing, however, does not rest on the discovery path alone: the same axis set is required to state the 101-fallacy decomposition (this paper), reappears in engineering failure analysis (oAOML), and its violation types are independently derivable from the conditions on coherent distinction-making (Paper III).

2.1 The Five Fundamental Boundary Violations

All logical fallacies, classical, modern, and rhetorical, can be classified as violations of one of five AOML constraint dimensions; §3.4 develops the corresponding analysis for fallacies as they arise in computational systems. These five types are exhaustive and non-overlapping at the structural level.

Type	Name	AOML Component	Core Failure
1	Axis Transfer Error	AXIS	One kind of truth substituted for another
2	Operator Overextension	OPERATOR	Illicit compositional rule applied (includes meta-axis substitution subtypes)
3	Verification Habitat Mismatch	METHOD	Wrong evidence type used for the claim's axis
4	Stratification Failure	LEVEL	Illegitimate level jump (includes TEMPORAL level-crossing)
5	Contrast Frame Collapse	Opposition topology (precondition)	False or inverted opposition imposed

Scope of the exhaustiveness claim. The claim is exhaustiveness over structural reasoning failure: cases in which the distinctions being drawn are themselves malformed, misassigned to a kind, composed without licensing, validated by the wrong instrument, moved across levels without warrant, or drawn within a mis-specified opposition space. It is not a claim over every way a reasoning process can arrive at a wrong answer. Two neighboring classes fall outside it. The first is quantitative failure: arithmetic error, measurement error, and miscalibrated credence, where every distinction is correctly drawn and correctly composed but the magnitudes attached to them are wrong. A forecaster with the right axis, operator, method, level and frame who is simply too confident has committed no structural violation; the framework has nothing to say about that error, and should not pretend otherwise. The second is performance failure: attention lapse, memory limitation, motivated reasoning considered as a cause rather than as a pattern. These explain why reasoners commit structural violations; they are not themselves violations of the conditions on distinction. Where a documented bias

has both a structural and a quantitative component, UFA classifies the structural component and is silent on the remainder.

Type 1: Axis Transfer Error. Occurs when a claim belonging to one semantic axis is treated as if it validated a claim on another axis. This substitutes truth-kind rather than merely evidence. Example (Ad Hominem): a claim about a speaker's SINCERITY is treated as evidence against the VERIDICAL truth of a proposition. Boundary violated: SINC → VER. Example (Naturalistic Fallacy): observed biological competition (VER) is used to justify ethical policy (NORM). Axis Transfer Errors are the most common and most rhetorically effective fallacies because they exploit genuine pattern correspondence while suppressing incompatible validation conditions.

Type 2: Operator Overextension. Occurs when a compositional operator is applied beyond its licensed semantic scope. Example (Post Hoc Ergo Propter Hoc): the CAUSE operator is illegitimately inferred from mere temporal succession or correlation. Example (Slippery Slope): the BECOME operator is extended into an unsupported multi-level causal chain without stratification control. AOML v2.2 introduces two important subtypes within Type 2:

Meta-Axis Substitution (R5, R6, R8, R9): a meta-axis modulation of a claim is treated as if it established the claim itself. This is the central result of AOML v2.2: a single schema, M[A] ↛ A, instantiated on each of the four meta-axes, generates a principled family of fallacies. DEG[A] → A (R5) yields Appeal to Probability, Appeal to Emotion, and Argumentum ad Populum (degree or intensity treated as truth); TEMP[A] → A (R6) yields Appeal to Novelty, Sunk Cost, and Presentism (temporal position treated as warrant); LIKE[A] → A (R8) yields Forgery, Sophistry, Hypocrisy, Demagoguery, Truthiness, and Illusion (resemblance to A treated as A); and EVID[A] → A (R9) yields the evidential appeals: hearsay-as-proof and inference-as-proof (the source-marking of a claim treated as establishing it). All four are Type 2 errors because the meta-axis modifier is an operator applied beyond its licensed scope; the derivative is not the function. That one substitution schema, ranging over four grammaticalized modulations, accounts for a large class of classical fallacies is the strongest single piece of evidence for the framework's predictive closure.

Polarity-Conditioned Overextension (R3, R7): CAUSE applied in its positive polarity to a target that licenses only negative causation. Example: "manufacture authentic goods" violates *CAUSE(***+) × AUTH, while "forge a document"** *(CAUSE(***−) × AUTH) is licensed. The polarity asymmetry is empirically documented in CLCA (Paper I, §4.4) and formalized as Structural Rule R3 in AOML v2.2. A second polarity-conditioned rule, R7 (agentive polarity asymmetry), is new in v2.2: the agentive ABSTR operator is licensed on the negative pole but blocked on the positive: languages lexicalize the agent of falsehood (a single word for liar) but provide no parallel agentive noun for the habitual truth-teller.* *The asymmetry is confirmed across both saturation sets; its mechanism remains open, though it aligns with the broader negative-asymmetry patterns in lexicalization documented by Horn (1989).***

Two further clarifications on Type 2 classifications. In cases like Circular Reasoning, the error lies not in the lack of logical movement, but in the illegitimate use of the IDENTITY operator to perform the function of an INFERENCE operator. IDENTITY is a relation of equivalence, not a derivation of truth; when it is used as a bridge between two distinct propositions to create the illusion of a result, it is an Operator Overextension. In cases like Amphiboly, the cause of the error is syntactic (grammatical ambiguity), but UFA classifies the semantic consequence: the ambiguity induces an unintended Operator Overextension in which ATTR or CAUSE is mapped to the wrong semantic constituents. UFA classifies what goes wrong in the reasoning, not what goes wrong in the grammar. The strength of the framework is that it stays in the realm of truth-claim composition rather than syntactic parsing.

Type 3: Verification Habitat Mismatch. Occurs when a claim is evaluated using an evidence type inappropriate to its semantic axis. The axis remains correct, but the validation method is illegitimate. Example (Appeal to Authority): expert testimony is treated as sufficient validation for a claim outside the expert's axis of competence. Example (Bandwagon): CONSENSUS is treated as evidence for VERIDICAL truth. AOML v2.2's validation habitats table (§IV) maps each axis to its default methods, admissible exceptions, and inadmissible methods, making Type 3 violations mechanically detectable. Note that "appeal to authority" names two structurally distinct errors: the Type 3 habitat mismatch described here (testimony used outside the expert's axis of competence) and a Type 2 evidential substitution (R9, EVID[A] $\nrightarrow$ A) in which the mere fact that a claim is authoritatively asserted or reported is treated as establishing it. The two readings are complementary, not competing.

Type 4: Stratification Failure. The primary form is stratification unassignability: no legal stratification assignment can make the judgment form well-formed. Example (Two-Truth Collapse): conventional truth is mistaken for ultimate truth, erasing level distinctions essential to semantic stability; restoring them is a habitat revision, not a premise repair. Level-crossing whose license can be repaired is classified Type 2 root with Type 4 locus and sits in the map's 2+4 compound family: Hasty Generalization (an empirical instance promoted to a category-level claim; the failed condition is the inductive license, and the instance/category transition is where the failure shows), and the TEMPORAL level-crossings introduced by AOML v2.2's TEMP meta-axis, such as Presentism and the Historian's Fallacy. The map's Second Edition exhibits the type's pure cases as unnumbered witnesses (the Liar cycle, the Yablo regress, Two-Truth Collapse), configurations that admit no well-founded level assignment at all (Paper III, §6).

The stratification distinction has deep roots. The four-level structure refines distinctions developed in classical Indian philosophy, particularly the Mādhyamaka Buddhist analysis of two truths articulated by Nāgārjuna in the second century. The two-truths doctrine, saṃvṛti-satya (conventional truth) and paramārtha-satya (ultimate truth), recognizes that claims operate at different levels of abstraction with different licensing conditions, and that confusing levels produces specific structural errors. AOML extends this analysis by adding EMPIRICAL and FORMAL as intermediate levels between concrete observation and ultimate metaphysical claims. The core insight, that

stratification levels constrain inference and that level-crossing without justification produces structural errors, is preserved from the Buddhist analysis. The history of how the two-truths doctrine has been translated into Western contexts illustrates the framework's analytical purchase on translation errors: the Sanskrit paramārtha-satya is variously rendered as "ultimate truth," "absolute truth," or "supreme truth" in English, each capturing part of the meaning while obscuring the level-distinguishing function. The two truths in their original articulation are not competing claims about the same reality at the same level but claims operating at structurally distinct levels. When translators flatten this into single-level evaluative vocabulary, the structural distinction collapses: a Type 4 stratification misalignment propagated through centuries of translation.

Type 5: Contrast Frame Collapse. Occurs when contrast is imposed where no true opposition exists, or when a multidimensional space is forced into a false binary. Example (False Dilemma): a non-binary domain is artificially reduced to mutually exclusive options. Example (Whataboutism): axis diversion combined with false equivalence, involving both Type 1 and Type 5 violations.

2.2 Mode-Condition Collapse: The Formal Mechanism

Definition: A Mode-Condition Collapse occurs when an analogical projection preserves Pattern-Correspondence while violating one or more required semantic constraints governing Manifestation-Mode, Verification-Method, Axis identity, or Stratification-Level as specified by the AOML tuple.

Let a claim be represented as an AOML tuple: (Axis, Operator, Method, Level). A valid analogical transfer from source S to target T requires: Pattern(S) ≅ Pattern(T); Axis(S) = Axis(T); Method(S) = Method(T); Level(S) preserved under inference; Operator(S) licensed in Axis(T). A Mode-Condition Collapse occurs whenever Pattern-Correspondence is preserved while one or more of these constraints is violated.

All five fundamental fallacy types are concrete instantiations of Mode-Condition Collapse: Type 1 (Axis condition violated), Type 2 (Operator licensing violated), Type 3 (Method condition violated), Type 4 (Level condition violated), Type 5 (Opposition topology violated). This formalism explains why fallacies are persuasive, recurrent, and cross-cultural: they preserve real patterns while collapsing the conditions that determine what kind of claim is being made and how it must be validated.

2.3 Structural Rules: Empirical Constraints from CLCA

AOML v2.2 introduces nine structural rules (R1–R9) that overlay the core legality matrix with empirical constraints discovered through CLCA. Each rule maps to one or more UFA violation types:

Rule	Name	Formal Statement	UFA Type
R1	Dispositional Causation Block	CAUSE × DISPOSITIONAL → universally blocked (16/16 languages)	Type 2
R2	Essence Stability	CAUSE × ESS → blocked; BECOME × ESS →	Type 2

Rule	Name	Formal Statement	UFA Type
		blocked	
R3	Authenticity Polarity Asymmetry	CAUSE(+) × AUTH → blocked; CAUSE(−) × AUTH → licensed	Type 2
R4	Positive Emergence Bias	BECOME(+state) >> BECOME(−state) across truth-domain axes	Type 2
R5	Degree Meta-Axis Substitution	DEG[A] ↛ A in any inference step	Type 2 (subtype)
R6	Temporal Meta-Axis Substitution	TEMP[A] ↛ A; TEMP level-crossing without justification	Type 2 / Type 4
R7	Agentive Polarity Asymmetry	ABSTR(agentive) × positive-pole → blocked; ABSTR(agentive) × negative-pole → licensed (liar lexicalizes; truth-teller does not)	Type 2
R8	Similative Meta-Axis Substitution	LIKE[A] ↛ A in any inference step	Type 2 (subtype)
R9	Evidential Meta-Axis Substitution	EVID[A] ↛ A in any inference step	Type 2 (subtype)

Prompt-layer disclosure. The rules in this table rest on two prompt architectures whose epistemic status differs, and the instruments are auditable in the public repositories. The naive pipeline's discovery prompts contain no axis inventory, no operator names, and no worked examples. The guided pipeline names territories and probe functions by design, and its v1.1 prompts carried three classes of leakage, disclosed here: a structural annotation on the degree territory that asserted the meta-axis analysis this paper defends; named probe functions, including revelation; and in-language worked example forms handed to five of the sixteen languages. All three were removed in deletion-only prompt variants (1.2 and 1.2.1), machine-verified so that every change is a deletion, with the full prompt diffs published alongside the raw outputs in the replication package. The de-leaked runs were pre-registered and executed on two model families. Three results bear on this table. The in-language forms are recovered without being given, five of five affected languages, both families. The revelation findings are reported with their full evidentiary history in the operator discussion above. And for R5, deleting the degree annotation removed the theory but not the pattern: truth-stem intensifier derivation re-emerges as a universal in both families, while no model spontaneously asserts the cross-cutting characterization itself. The data pattern belongs to the corpus; the meta-axis analysis remains, explicitly, the analysts' contribution on top of it.

3. The Calculable Space of Error: Structure and Evidence

If the AOML constraint geometry is correct, then the space of possible reasoning errors is not open-ended. It is bounded by the geometry itself. The classification is not a historical inventory of named errors; it is a structural projection of the AOML constraint space. With the catalog specified, the possible violation sites can be enumerated from

the geometry itself: the illegal cells of the legality matrix, the inadmissible method-axis pairings of the habitats table, the four meta-axis substitutions, the level-crossing pairs. The named fallacies of the tradition occupy positions in that space rather than define it, each at a determinate position fixed by the AOML component violated and the specific Mode-Condition Collapse that occurs; positions can exist before their occupants are named, as the evidential substitution family (R9) did.

Because AOML fully specifies the dimensions required for semantic coherence, the table is closed under extension: no stable fallacy can exist outside its structure. Newly identified reasoning errors must either instantiate a single violation or decompose into a compound of them. The table functions as a Universal Fallacy Predictor, not merely a classifier. This is the principal claim of UFA, and it is falsifiable.

3.1 Representative Fallacy Mappings

Fallacy	Type	Collapsed Condition	Structural Description
Ad Hominem	1	Axis Identity	SINC → VER: speaker character substituted for propositional truth
Naturalistic Fallacy	1	Axis Identity	VER → NORM: descriptive fact treated as normative justification
Post Hoc Ergo Propter Hoc	2	Operator Licensing	CAUSE inferred from temporal succession or correlation
Slippery Slope	2	Operator Licensing	BECOME extended into unsupported multi-level causal chain
Appeal to Probability	2	Meta-Axis Substitution	DEG[VER] → VER: high likelihood treated as established truth (R5)
Appeal to Novelty	2/4	Meta-Axis + Level	TEMP[NORM] → NORM: recency as normative validity (R6)
Sunk Cost	2/4	Meta-Axis + Level	∫past → future obligation: temporal investment treated as future warrant (R6)
Bandwagon	3	Validation Habitat	CONSENSUS substituted for OBSERVATION/EXPERIMENT on VER axis
Appeal to Authority	3	Validation Habitat	TESTIMONY from wrong axis treated as evidence on target axis
Hasty Generalization	4	Stratification	EMPIRICAL instance → CATEGORY claim without justification
Presentism	4	Temporal Level	Present-level standards applied to past-level events (R6)
False Dilemma	5	Opposition Topology	Multidimensional domain forced into binary opposition
False Equivalence	5	Opposition Topology	Superficial similarity treated as structural identity

Fallacy	Type	Collapsed Condition	Structural Description
Whataboutism	1+5	Axis + Contrast	Axis diversion (SINC → VER) via illegitimate contrast
Forgery	2	Meta-Axis Substitution	LIKE[AUTH] → AUTH: a copy passed off as genuine (R8)
Hearsay as Proof	2	Meta-Axis Substitution	EVID[VER] → VER: a reported source treated as establishing the claim (R9)

Note: The complete mapping of 101 fallacies to UFA types is provided in the companion Fallacy Map document (Second Edition, https://doi.org/10.5281/zenodo.21880635). The entries above are representative, selected to illustrate each type and the subtypes introduced by AOML v2.2 (meta-axis substitution across all four meta-axes, polarity conditioning, temporal level-crossing).

3.2 Structural Interpretation

Each row in the table represents a lawful failure mode of analogical reasoning: the pattern holds (why the fallacy persuades) and the conditions collapse (why it is invalid). The table shows that fallacies differ not by rhetorical style but by which semantic constraint fails. This explains why very different-looking arguments can commit the same fallacy, and why the same fallacy appears across cultures, disciplines, and historical periods.

3.3 Exhaustiveness and Predictive Closure

Because AOML enumerates all dimensions required for semantic legality, the space of structural error is calculably closed: no sixth primitive fallacy type is possible without introducing a new semantic dimension. Apparent new fallacies must reduce to a single Mode-Condition Collapse or a compound of known collapses. This property distinguishes the framework from traditional logic taxonomies and enables direct architectural enforcement in artificial intelligence systems.

De Morgan doubted that the ways by which human beings arrive at error could ever be classified (De Morgan, 1847). UFA does not dispute this. It distinguishes the potentially unbounded causal and rhetorical paths to error from the structural dimensions violated by an invalid composition. Its claim to universality concerns the latter, not the former.

An important distinction: the exhaustiveness claim has two levels. The first is definitional: given AOML's five constraint dimensions, violations must fall into one of five types. The second is empirical: that these five types actually account for all known structural reasoning errors (§2.1). The companion Fallacy Map document demonstrates this empirically by mapping 101 classical and contemporary fallacies to the five types. Every fallacy in the inventory decomposes cleanly. None requires a sixth type. The map's Second Edition strengthens this test: every entry was re-examined under a classification criterion frozen before the pass, which relocated roughly a fifth of the entries, demonstrating the procedure's power to move them; recovered two established

fallacies whose structures had been bundled inside other entries (Aggregation Reversal and Confusion of the Inverse, map #102 and #103); and left no entry outside the five types. A taxonomy sorts the entries it is given; an architecture can show that the entries it was given were the wrong ones. That is the evidence for exhaustiveness, not a mathematical proof but a comprehensive empirical test. If a sixth type were discovered, it would mean AOML is missing a constraint dimension, and the primitive set would need expansion. But the five types already identified would remain valid; a sixth would be an addition, not a negation. The framework is designed to be extensible in this way. The falsification target should therefore be stated precisely: what would refute the architecture is not the discovery of an unfamiliar fallacy name but a reasoning error that resists decomposition into the five types under the AOML constraint dimensions, or a revision to the CLCA substrate that removes the constraint a type is derived from. Extensibility protects the inventory; it does not protect decomposability. Nor do structure-level failures escape the types: configurations whose every local step is legal but whose dependency structure is ungrounded (circular justification, Yablo-type infinite sequences) are violations of the existing conditions read globally over the argument's dependency structure, not a sixth type (Paper III, §6).

AOML v2.2's meta-axis substitution family (R5, R6, R8, R9) does not introduce new primitive types; it reveals a single subtype within Type 2 (Operator Overextension), the substitution schema M[A] $\nrightarrow$ A, that unifies a large class of previously unclassified or misclassified fallacies. Similarly, the polarity conditioning on CAUSE (R3) and the agentive polarity asymmetry on ABSTR (R7) do not add new types but refine the licensing conditions under which Type 2 violations occur. The expansion from two meta-axes to four leaves the five-type taxonomy untouched, every new rule lands inside Type 2, which is itself evidence for the closure claim.

3.4 Implications for AI and Translation

In computational systems, each fallacy corresponds to an illegal AOML state transition. Enforcing AOML constraints prevents fallacies before they are generated, rather than detecting them post hoc. Catastrophic translation errors arise when semantic axes, methods, or levels are collapsed during cross-linguistic mapping, demonstrating that logical fallacies and translation failures share a common structural cause.

3.5 Predictive Power and Falsifiability

Prediction 1: Exhaustiveness. No stable reasoning error can exist outside the five fundamental boundary violations. The discovery of a persistent, cross-contextual fallacy that cannot be decomposed into any of the five types would falsify UFA or require revision of the AOML primitives.

Prediction 2: Decomposability. Newly named or domain-specific fallacies will reduce to single or compound Mode-Condition Collapses. For any proposed fallacy F, there exists a mapping F $\rightarrow$ {C_1, …, C_n} where each C_i is a violation of one AOML constraint dimension.

Prediction 3: Translation Isomorphism. Catastrophic cross-linguistic translation errors will be structurally isomorphic to logical fallacies. Example: the collapse of the Sanskrit distinction between saṃvṛti-satya (conventional truth) and paramārtha-satya (ultimate truth) should map to Type 4 Stratification Failure; the map's Second Edition exhibits exactly this configuration as its witness W3 (Two-Truth Collapse).

Prediction 4: Architectural Prevention in AI. Systems enforcing AOML constraints at generation-time will exhibit lower rates of category error, reduced hallucination involving evidence laundering, and improved consistency across domains, and these improvements should be structurally distinct, not merely statistical.

Prediction 5: Boundary Sensitivity. Robust intelligence, human or artificial, will demonstrate boundary sensitivity: the ability to detect when a pattern transfer violates its validation conditions and to halt or reframe reasoning accordingly.

3.6 Compound Violations

Several reasoning errors involve violations of multiple types simultaneously. The framework analyzes these as compound violations rather than as additional types. Whataboutism combines Type 1 (axis diversion from VER to SINC) with Type 5 (false equivalence between two SINC claims). Appeal to Tradition combines Type 3 (CONSENSUS as validation for NORM) with Type 2 (TEMP meta-axis substitution). Presentism combines Type 4 (level misalignment between past and present) with Type 2 (TEMP meta-axis substitution). Tu Quoque combines Type 1 (SINC → VER substitution) with Type 5 (false equivalence implying mutual cancellation). Compound violations are not a separate category; they are reasoning errors where multiple structural violations co-occur, each analyzable under the framework's vocabulary. The compound nature is itself analytical content: recognizing that a fallacy combines axis substitution with contrast frame collapse explains why the fallacy is more rhetorically effective than either violation alone, because it exploits two independent failure modes simultaneously.

3.7 Boundary Sensitivity and Early Error Arrest

Boundary sensitivity is the capacity to detect when an inference, analogy, or translation step violates semantic constraints before the violation propagates. In humans, this is evident in humor, irony, and cognitive dissonance, phenomena revealing that meaning preservation depends on the ability to recognize when a pattern transfer has exceeded its validation conditions.

In artificial systems, the absence of boundary sensitivity manifests as confident persistence in structurally invalid states, particularly when early semantic errors are reinforced through optimization or training. UFA predicts that boundary sensitivity is not optional but diagnostic of robust intelligence.

4. Architectural Integration: From Fallacy Prediction to Prevention

4.1 UFA as a Semantic Legality Layer

Classification is useful, but it is not the point. The point is prevention. If you know the complete set of ways a program can crash, you can build a runtime that prevents those crashes before they happen. That is what UFA provides for reasoning systems. A verified reference implementation of this architecture exists, and its existence and role are on the public record in the engineering companion's account of the oAOML program as three artifacts: the evidence ledger, the enforcement implementation, and the sealed preregistered predictive test (oAOML: Engineering Failure Analyses, §6, https://doi.org/10.5281/zenodo.21960566).

In the implementation series, UFA functions as a semantic legality layer governing permissible state transitions during reasoning and generation. Every candidate claim is represented as an AOML tuple (Axis, Operator, Method, Level). Proposed inference steps are evaluated as state transitions in this space. Transitions that violate AOML constraints are blocked or flagged before continuation. This prevents fallacies as illegal moves, not as detected mistakes.

4.2 The Negative Space in Training Data

Language models learn from human-produced text. That text reflects the errors humans actually produce, but it also reflects, less obviously, the absence of errors humans catch before producing them. When a competent reasoner considers asserting something that crosses a semantic boundary, the assertion often does not reach the page. This creates an asymmetry in training signal: models learn the distribution of human output, including its fallacies, but they do not learn the much larger distribution of errors humans consider and reject. The structurally possible violations that humans rarely make, because cognitive architecture and linguistic convention block them, are the violations least represented in training data, and therefore the violations models are least equipped to detect or avoid. UFA addresses this directly. The framework characterizes not just what humans do produce but what languages systematically block from production: the dispositional causation that no language permits morphologically, the polarity asymmetries that license forgery while blocking the manufacture of authenticity, the structural gaps that appear consistently across unrelated language families. A language model has no such architecture; it inherits only the statistics of human output. Explicit characterization of what languages refuse, the negative space, provides information that corpus statistics cannot recover.

4.3 Prevention of Hallucination and Evidence Laundering

Many AI hallucinations are not fabrications of facts but something more subtle: evidence laundering. The system takes a validation method that is legitimate on one axis and quietly applies it to a claim on a different axis. A consensus statistic becomes veridical evidence. Resemblance to truth (LIKE[VER]) gets mistaken for factual accuracy (VER), the meta-axis substitution prohibited by Rule R8. Testimony from one domain becomes

proof in another. These are not random errors; they are Type 1, Type 2, and Type 3 violations, and they are structurally predictable.

Under UFA enforcement, axis identity must be preserved across inference chains, validation methods must remain within their licensed habitats, and level transitions require explicit justification. This directly addresses failure modes observed in large language models that appear fluent yet semantically unstable.

A useful distinction: not all hallucinations are the same kind of error. Reasoning-error hallucinations involve invalid inference structures and are caught by operator-licensing checks. Confabulation involves filling gaps with structurally plausible but unsupported content and is caught by method-licensing checks. Factual-error hallucinations involve incorrect claims at the content level and are not directly addressed by structural checks, since the structural form may be correct even when the factual content is wrong. UFA enforcement targets the first two categories; the third requires content verification beyond structural analysis. Evaluation that distinguishes these categories provides a sharper test of the framework's effects than aggregate hallucination metrics.

A note on training methodology: the natural response to violations in training data is to filter them out. This is a weaker intervention than the alternative. Filtered training data leaves gaps; the model never encounters certain patterns and may fail to recognize them when users present them. Labeled training data preserves the examples with structural annotation: the model encounters the violation and the fact that it is a violation, simultaneously. This enables a capability that current models struggle to maintain: representing a claim without committing to it. A model that has learned to distinguish content from commitment can discuss fallacious reasoning without reproducing it as valid, can represent what a speaker said without treating it as established. The difference between filtering and labeling is the difference between making a model ignorant of fallacies and making it competent about them.

4.4 Translation and Cross-Cultural Reasoning

Because UFA operates at the level of semantic structure rather than surface form, it provides a shared constraint space for translation and cross-cultural reasoning. Translation is treated as AOML-preserving mapping, not lexical substitution. Semantic axes and levels must be conserved across languages. Violations manifest as the same Mode-Condition Collapses identified in §2 and §3.

4.5 Structural Detection of Manipulation

The dispositional causation block (R1) has an unexpected implication for understanding manipulation. Manipulation that targets sincerity, authenticity, or genuine conviction is attempting an operation that CLCA shows cannot achieve its stated goal directly. Forced sincerity is not sincerity; manufactured conviction is not conviction; coerced loyalty is not loyalty. The manipulator must therefore proceed indirectly, simulating the result through coercion, social pressure, and information control. This is why manipulation always involves concealment: you cannot openly cause a dispositional

state, so you must create conditions that simulate it while hiding the construction. The structural impossibility of direct causation forces manipulation into deceptive form.

This explains why propaganda fails to produce genuine conviction even when it succeeds at producing behavioral compliance. The gap between coerced behavior and authentic inner state is documented across political theory (Havel, Arendt) and is structurally necessary rather than contingent: direct causation of dispositional states is incoherent, so coercion can only ever produce the behavioral surface, not the inner state it simulates. For computational systems, detection of manipulation becomes detection of the specific UFA violation patterns that propaganda exploits, regardless of which political position they serve. A UFA-aware analysis flags binary collapse, axis substitution, or manufactured consensus as structural features whether they appear in any form of rhetoric. The detection is structural, not partisan.

4.6 Relation to Reward and Learning Systems

UFA is orthogonal to reward optimization. It does not assign value or preference; it enforces semantic legality. Learning systems may explore freely within the legality space. Rewards shape behavior, but constraints shape possibility. This separation prevents reward hacking via semantic collapse.

4.7 Temporal Trojan Horses in Translation and Training Pipelines

Semantic boundary violations are not temporally neutral. When a Mode-Condition Collapse occurs early in a reasoning or training pipeline, particularly during translation, annotation, or curriculum construction, it can function as a Temporal Trojan Horse: a structurally invalid assumption that is subsequently treated as legitimate input and propagated forward.

In artificial intelligence systems, this effect is amplified by training dynamics. Early misaligned labels, mistranslations, or category confusions are reinforced through optimization, causing the system to internalize structurally invalid inference patterns as stable associations. Because learning systems optimize for internal coherence and predictive success rather than semantic legality, these violations can persist and spread even as surface accuracy improves.

The Universal Fallacy Architecture provides a principled response: by enforcing AOML constraints at the point of data ingestion and transformation, systems can prevent the introduction of Temporal Trojan Horses altogether. This shifts error handling from retrospective correction to architectural inoculation.

A note on the level of description: the Temporal Trojan Horse concept is defined here at the semantic level, in terms of AOML constraint violations propagating through a pipeline. The mechanism by which such violations become structurally embedded during gradient descent in transformer architectures, specifically how an invalid semantic invariant gets reinforced through optimization rather than corrected, is an implementation question that belongs to the implementation series, where the runtime enforcement work (Belief Provenance Graph) and training-time protection work (Reflective Constraint Barrier) address how to detect and prevent these violations within

specific computational architectures. This paper establishes what the violations are and why they propagate.

5. Deeper Hypothesis: Target-Type Classes and the Unity of CLCA Findings

***Status: This section presents a structural hypothesis derivable from the CLCA data but not yet independently confirmed. It is offered as a candidate deeper explanation, not as an established result. The five UFA types and their structural rules stand on their own empirical and analytical grounding; the hypothesis developed here proposes a possibly-deeper layer that, if correct, would explain why the rules cluster as they do.*

5.1 The Pattern in the Data

Three of the strongest empirical findings from CLCA share a feature that has not yet been given a unified structural account. The dispositional causation block (R1) prohibits direct external causation of character states across all sixteen languages. The positive emergence bias (R4) favors BECOME(+state) over BECOME(−state) across truth-domain axes. The REVEAL operator preference treats certain semantic targets as latent and pre-existing rather than constructed, running world-to-claim rather than claim-to-world. Each finding has been treated as a separate empirical regularity governed by its own structural rule. But examined together, they share a common feature: each concerns how languages treat the relationship between an external agent and a target state. R1 says external agents cannot directly cause certain target states. R4 says certain target states emerge naturally rather than being externally constructed. REVEAL says certain target states are discovered rather than made. This is not three separate findings about three separate phenomena. It is one finding about a structural property of the targets themselves.

5.2 The Target-Type Hypothesis

The hypothesis: languages distinguish between target types based on how their constitutive structure relates to external causation. This distinction is structurally prior to the operator licensing matrix and explains why specific operators are licensed or blocked on specific targets. Three target-type classes emerge from the data. FACTUAL targets are states of the external world that license external causation; physical facts can be caused, changed, or destroyed by external agents, and the state is constituted independently of how it came to be. EPISTEMIC targets are knowledge or belief states that prefer discovery and revelation over external construction; truth-as-known, evidence, and recognition are discovered through investigation rather than manufactured. DISPOSITIONAL targets are character states or inner states that universally block direct external causation and require internal cultivation; sincerity, honesty, authentic conviction are constituted partly by their non-coerced origin, which is what makes the operation incoherent rather than merely difficult. Under this hypothesis, R1 blocks CAUSE on DISPOSITIONAL because dispositional states cannot be externally caused. R4 favors positive emergence because emergence is the licensed

mode for non-FACTUAL targets. The REVEAL preference reflects the EPISTEMIC and DISPOSITIONAL bias toward discovery over construction. Three findings reduce to one. The hypothesis is independently anticipated by the causative-typology literature: morphological and lexical causatives cross-linguistically encode direct causation (Shibatani 1976; Dixon 2000), causal semantics distinguishes direct from mediated force transmission (Wolff 2003), and dispositional states resist direct external causation for constitutive reasons. On this reading, the character-trait causation block is the intersection of a known directness constraint with the DISPOSITIONAL target class: convergent human-derived typology arriving at the same layer from independent evidence. One caveat the convergence sharpens: the current CLCA probe sets test agentive-direct causation but do not systematically separate agentive from circumstantial or experiential causes ("the war made him cynical"), which the directness account predicts should be acceptable; the agentive/circumstantial split is a designed test of the native-speaker validation round (Paper I, Appendix B).

5.3 Why This Would Matter

If the target-type hypothesis is correct, the implications are substantial. The framework's analytical depth increases: the five UFA types remain valid as descriptions of where Mode-Condition Collapse occurs, but their derivation becomes deeper. Many fallacies currently classified as Type 1 (axis substitution) might be better understood as target-type crossing errors, with axis substitution being the surface manifestation. Think of it as the difference between cataloguing which specific instructions crash the processor versus understanding why the processor architecture makes certain instruction classes illegal in the first place. The framework's predictions also become more testable: languages should treat all DISPOSITIONAL targets the same way regardless of which axis they fall on. Trust, faith, conviction, love, respect should all pattern with sincerity in resisting direct causation. For AI safety, the implication extends directly: if certain target types resist direct causation as a matter of structural coherence, then AI systems optimized to produce sincerity-signals or authenticity-signals are attempting operations that cannot succeed in the form they appear to attempt. The simulation of these states bypasses their constitutive conditions.

5.4 What Would Confirm or Refute the Hypothesis

The hypothesis makes specific predictions that systematic CLCA work could test. First: CAUSE will be universally blocked on additional dispositional targets beyond sincerity and authenticity. Trust, faith, conviction, love, respect, devotion, and similar inner states should pattern with sincerity in resisting direct causation across the language sample. If they do, the target-type hypothesis is supported; if some allow direct causation while others block it, the hypothesis needs refinement. Second: REVEAL will be preferred over CAUSE on additional epistemic targets. Knowledge claims, recognition claims, understanding claims should pattern with truth-discovery in preferring revelation over construction. Third: the BECOME(+) bias will appear specifically for non-FACTUAL targets. Factual states should not show the asymmetry as strongly because factual states are equally producible and destroyable through external causation. Fourth: target-type membership should predict cross-linguistic operator-licensing patterns more

accurately than axis-specific stipulations alone. These predictions are testable through systematic CLCA expansion to the relevant target types. The work is bounded: it requires extending existing methodology to additional concepts rather than developing new methodology.

5.5 Relation to the Five UFA Types

The hypothesis does not displace the five UFA types. Even if target-type structure is real, the five types remain valid as descriptions of where Mode-Condition Collapse occurs in compositional reasoning. What changes, if the hypothesis is confirmed, is the explanatory depth. The five types describe what fallacies look like; the target-type structure would explain why specific operators are blocked on specific targets in the first place. The relationship is similar to that between observed regularities in chemistry and the underlying atomic theory that explains them: the regularities were valid before atomic theory; atomic theory explained why the regularities held the form they did. For practical applications, the five types are sufficient. AI systems enforcing AOML constraints can do so based on the operator-licensing matrix without needing to resolve whether the matrix follows from target-type structure or from independent axis-specific rules. For the broader research program, the hypothesis matters because it points toward the deeper architecture that CLCA may be partially recovering. The research program's eventual mature form may organize itself around target types as primary primitives with axes and operators as derived structure. This is a hypothesis the work might eventually justify; it is not a claim the work currently establishes.

6. Limitations, Scope, and Future Work

6.1 Scope of the Current Framework

UFA is derived from CLCA findings focused primarily on truth-related semantic domains. While foundational, the present formalism does not yet claim exhaustive coverage of temporal reasoning beyond truth predicates, spatial reasoning and embodiment, emotional and affective inference, or pragmatic speech acts outside propositional structure. Extension to these domains is a matter of empirical expansion, not theoretical assumption.

6.2 Linguistic and Cultural Coverage

CLCA results supporting AOML and UFA draw from sixteen languages across ten families and three isolates. While convergence across unrelated families strongly suggests non-arbitrariness, universality claims remain provisional pending broader inclusion of indigenous and under-represented language families, systematic testing in sign languages and non-spoken modalities, and independent replication by researchers not aligned with CIF development.

6.3 Cognitive and Neurobiological Claims (Explicitly Not Made)

UFA does not claim that AOML primitives are consciously represented, that they correspond directly to neural modules, or that humans reason by explicit tuple evaluation. The framework is structural and functional, not reductionist. Its validity rests on whether semantic collapse occurs reliably when constraints are violated, not on claims about mental implementation.

6.4 AI Implementation Limits

While UFA enables architectural prevention of known failure modes, several limits should be stated clearly. UFA does not guarantee truth: a claim can be structurally well-formed and factually wrong, since the framework enforces compositional legality, not correspondence with reality. Factual-error hallucinations are not directly addressed by structural enforcement. UFA does not eliminate adversarial misuse: a system enforcing AOML constraints can still be used for unintended purposes, and adversarial actors can construct AOML-legal arguments that serve manipulative purposes when the structural moves themselves are not the primary site of manipulation. UFA does not replace epistemic judgment or moral reasoning: it constrains how claims may be made, not which claims are worth making. A system producing only AOML-legal outputs may still produce outputs that are unhelpful, incomplete, or normatively problematic. UFA enforcement introduces computational overhead: tagging text with AOML tuples, evaluating transitions, and providing structured feedback each requires resources that current systems do not allocate to structural analysis. Finally, the proposals developed in §4 for labeling-not-filtering training data, structured feedback, negative-space characterization, and structural manipulation detection are theoretical contributions whose practical implementation depends on engineering work beyond this paper. Whether these proposals produce their predicted effects is an empirical question that requires building and testing the systems described. Preliminary implementation results, in preparation as a separate engineering series, are consistent with Predictions 3 and 4; the present paper's claims do not rest on them.

6.5 Epistemic Status of CLCA Findings

Paper I presents the CLCA discoveries as candidate regularities requiring continued validation, not as definitive universals. UFA inherits this epistemic status: the framework’s predictive power depends on the stability of the underlying constraint geometry. If future CLCA expansions reveal that certain constraints are less universal than the current 16-language sample suggests, the corresponding UFA predictions would require revision. The framework is designed to be falsifiable at this level.

6.6 Relationship to Existing Philosophical and Logical Traditions

The framework’s relationship to longstanding philosophical and logical traditions deserves more systematic treatment than this paper provides. Initial observations have been integrated where they do specific analytical work: the Buddhist analysis of two truths in the Type 4 discussion (§2.1), the connection to Hume on the is-ought distinction through the Naturalistic Fallacy, Ryle on category errors through Type 1

analyses, and the convergence of the four primary violation types with traditions in formal logic, informal logic, scholastic philosophy, and Buddhist stratification analysis. These connections suggest that the framework formalizes structural intuitions that multiple intellectual traditions have articulated independently, which is consistent with the framework capturing real features of reasoning rather than imposing arbitrary categories. Systematic engagement with these traditions, including the analytic philosophy of language, phenomenological analysis of authentic inner states, Buddhist philosophy of stratification, and scholastic logic, would strengthen the framework's intellectual standing and likely refine specific aspects of the formalization. This engagement is reserved for future work rather than attempted here.

Within fallacy theory proper, the framework's nearest relatives are the modern catalogue tradition and the argumentation-theoretic tradition. Hamblin (1970) reopened the systematic study of fallacies; Walton's argumentation schemes (1995) give many fallacies a structural, condition-based analysis; pragma-dialectics (van Eemeren & Grootendorst, 2004) treats fallacies as violations of the rules of a critical discussion; and Toulmin's (1958) field-dependence of argument anticipates the validation-habitat idea (Type 3) directly. What UFA adds is orthogonal to each. The five types are not catalogued from usage but derived from cross-linguistically attested compositional constraints (Paper I), and their exhaustiveness is argued structurally rather than asserted (Paper III). Where Walton's schemes and pragma-dialectics' rules are normative-procedural, what a good arguer or a well-run discussion should do, UFA is semantic-structural: which compositions the geometry of truth-claims licenses, independent of any discussion protocol. The contribution is the derivation and the closure, not the recognition that fallacies have structure, which these predecessors established. Appendix A tabulates this comparison: the organizing principles of the three traditions side by side, and a set of representative fallacies classified under each. A companion Terminology Correspondence table maps the framework's vocabulary, including the five violation types, to its nearest counterparts in argumentation theory, philosophy of logic, cognitive science, and AI, with each correspondence graded rather than asserted as an equivalence (https://doi.org/10.5281/zenodo.21781768).

6.7 Future Directions

One methodological commitment has been registered and its first phase executed: an inter-rater reliability study of the five-type classification, in which independent raters receive only the type definitions (with every named-fallacy example removed) and classify the map's 94 entries (the inventory as of the study; the map's Second Edition now describes 101 distinct fallacies) alongside twenty salted distractor items (valid argument forms and non-fallacious reasoning flaws), with not-a-fallacy and fits-no-type options available throughout; thresholds and the interpretation of intermediate outcomes were committed before any rating. In the pre-registered first phase, three heterogeneous frontier language models rated all items blind. The falsifiability results were maximal: raters rejected 30 of 30 valid-argument distractors and 30 of 30 non-fallacious-flaw distractors, and applied the fits-no-type option to 14.9 percent of genuine inventory items. The classification procedure demonstrably recognizes items that do not fit, which is the substance of any unfalsifiability-in-practice concern about Prediction 2;

because the distractors were pattern-level items with informative names, these rejection rates are an upper bound, and the second phase uses instance-level argument texts with neutral wording. Agreement on type assignment was moderate (Fleiss's kappa of 0.51, with majority agreement on 87 of 94 items), landing in the pre-registered intermediate band whose interpretation was fixed in advance: the five types carve coarse structure reliably, but example-free definitions underdetermine boundary assignment, with disagreement concentrated in the meta-axis substitution family that the anti-leakage redaction had deliberately stripped of its examples. Two refinements are pre-declared for the second phase: human raters, and a definitions sheet with synthetic examples authored independently of the inventory; if agreement fails to improve under the latter, that is evidence the boundary problem lies in the taxonomy rather than the materials, and it will be reported as such. The reported statistics stand as computed against the First Edition classification key; the pre-declared second phase runs against the Second Edition definitions and key, since the Type 4 definition changed intensionally in the Second Edition. The frozen protocol, materials, and raw ratings are archived under the study's registration hash. Beyond this, key directions include: cross-domain CLCA beyond truth semantics, examining whether AOML-like compositional structures appear in other semantic territories; native-speaker validation of the cross-linguistic findings underlying the structural rules, as described in Paper I’s Appendix B; empirical testing of the target-type hypothesis (§5), particularly the predictions that CAUSE will be universally blocked on additional dispositional targets and that REVEAL will be preferred on additional epistemic targets; engineering development of AOML-aware AI systems, including tagger development, labeled-corpus training experiments, and structured-feedback fine-tuning approaches; quantitative benchmarks comparing constrained against unconstrained AI on tasks where the framework predicts specific failure modes; developmental and psycholinguistic studies testing boundary sensitivity; long-horizon stability tests for self-modifying systems under UFA constraints; and systematic engagement with the philosophical traditions identified in §6.6. These directions are independent of one another: progress on any one does not require progress on the others, and the framework’s contributions can be strengthened along multiple axes simultaneously by different research programs working with different methodologies.

6.8 Use of AI

The cross-linguistic evidence this paper builds on was elicited using large language models as structured linguistic informants, as documented in Paper I (§2.5); this paper inherits that instrument and its limitations. Within this paper's own work, the executed first phase of the inter-rater reliability study (§6.7) used three heterogeneous frontier language models as blind raters, as disclosed there; the pre-declared second phase uses human raters.

The author also used an AI assistant for manuscript preparation, including copy-editing, formatting, and drafting assistance on prose that was subsequently reviewed and revised by the author. All scholarly content, arguments, and final wording are the author's responsibility.

7. Conclusion

The Universal Fallacy Architecture establishes that reasoning errors, translation failures, and AI hallucinations are not unrelated phenomena, but expressions of a single underlying fact: meaning collapses predictably when its structural constraints are violated.

Built on the three-layer AOML v2.2 architecture derived from CLCA, UFA identifies five fundamental boundary violation types that are exhaustive and predictive. The meta-axis substitution family (R5, R6, R8, R9: one schema, M[A] ↛ A, over four grammaticalized modulations), polarity-conditioned operator licensing (R3, R7), and the target-type hypothesis provide a richer and more empirically grounded framework than the original formulation, while maintaining the same structural parsimony: five types, one mechanism (Mode-Condition Collapse), one constraint geometry (AOML).

The framework is falsifiable, empirically testable, and independent of philosophical allegiance. It does not claim that humans consciously apply AOML constraints, only that meaning collapses reliably when they are violated. This makes UFA suitable as both a diagnostic theory of reasoning error and a prescriptive architecture for artificial intelligence systems.

Paper I recovered the instruction set. This paper catalogues the illegal operations. Paper III establishes why there are exactly five types of illegal operation, grounding the structural argument in the conditions for coherent distinction-making itself. Paper IV shows what truth is on each axis when the conditions are satisfied, resolving the philosophy of truth debate as a Type 1 pattern in which each tradition correctly identified one axis's truth-structure and then overgeneralized. Together, these four papers constitute the theoretical foundation of the Convergent Semantic Architecture: what the constraints are, what happens when they are violated, why the violation space takes the form it does, and what truth consists of on each face of the geometry. Implementation, including runtime enforcement, training-time protection, and the broader engineering program, builds on this foundation as a subsequent body of work. The question that started this research was practical: how do you build AI systems that handle meaning correctly? The answer turns out to require understanding how meaning works in the first place. That understanding is what CLCA provides, what AOML formalizes, what UFA makes enforceable, and what the axis-relative theory of truth grounds in the deepest structural terms available.

References

Seeley, B. A. (2026). Cross-Linguistic Compositional Analysis (CLCA): A Reproducible Protocol for Recovering Shared Conceptual Structure from Cross-Linguistic Evidence. Paper I, Foundations of the Convergent Semantic Architecture. https://doi.org/10.5281/zenodo.21709636

Seeley, B. A. (2026). AOML v2.2: Semantic Constraint Architecture. Companion document to Paper II. https://doi.org/10.5281/zenodo.21792514

Seeley, B. A. (2026). The Universal Fallacy Map (Second Edition). Companion document to Paper II. https://doi.org/10.5281/zenodo.21880635

Seeley, B. A. (2026). oAOML: Engineering Failure Analyses. Canonical Incidents Mapped to the Five Structural Violation Types. Companion document to Papers II and III. https://doi.org/10.5281/zenodo.21960566

Seeley, B. A. (2026). Terminology Correspondence: Mapping CLCA/AOML/UFA Vocabulary to Standard Disciplinary Terms. Companion document to the Foundations of the Convergent Semantic Architecture series. https://doi.org/10.5281/zenodo.21781768

Aristotle. (1984). On Sophistical Refutations (W. A. Pickard-Cambridge, Trans.). In J. Barnes (Ed.), The Complete Works of Aristotle: The Revised Oxford Translation (Vol. 1, pp. 278–314). Princeton University Press. (Original work c. 350 BCE, 164a20–184b8)

De Morgan, A. (1847). Formal Logic: or, The Calculus of Inference, Necessary and Probable. Taylor and Walton.

Dixon, R. M. W. (2000). A typology of causatives: form, syntax and meaning. In R. M. W. Dixon & A. Y. Aikhenvald (Eds.), Changing Valency: Case Studies in Transitivity (pp. 30–83). Cambridge University Press.

Hamblin, C. L. (1970). Fallacies. Methuen.

Horn, L. R. (1989). A Natural History of Negation. University of Chicago Press.

Hume, D. (1740). A Treatise of Human Nature, Book III: Of Morals. Book III, Part I, Section I (T 3.1.1.27; SBN 469–470 in the Selby-Bigge/Nidditch edition, Clarendon Press, 1978).

Ryle, G. (1949). The Concept of Mind. Hutchinson.

Shibatani, M. (1976). The grammar of causative constructions: A conspectus. In M. Shibatani (Ed.), Syntax and Semantics 6: The Grammar of Causative Constructions (pp. 1–40). Academic Press.

Sweetser, E. (1990). From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure. Cambridge University Press.

Toulmin, S. E. (1958). The Uses of Argument. Cambridge University Press.

van Eemeren, F. H., & Grootendorst, R. (2004). A Systematic Theory of Argumentation: The Pragma-Dialectical Approach. Cambridge University Press.

Walton, D. (1995). A Pragmatic Theory of Fallacy. University of Alabama Press.

Walton, D., Reed, C., & Macagno, F. (2008). Argumentation Schemes. Cambridge University Press.

Wolff, P. (2003). Direct causation in the linguistic coding and individuation of causal events. Cognition, 88(1), 1–48. https://doi.org/10.1016/S0010-0277(03)00004-0

Yablo, S. (1993). Paradox without self-reference. Analysis, 53(4), 251–252. https://doi.org/10.1093/analys/53.4.251

Note: The complete 101-fallacy mapping table is provided in the companion Fallacy Map document (Second Edition, https://doi.org/10.5281/zenodo.21880635) rather than reproduced in full here.

Appendix A: Comparison with Traditional Taxonomies

This appendix summarizes the comparison developed in §6.6 in tabular form. Table A1 contrasts the organizing principles of the Aristotelian catalogue tradition, Walton's argumentation schemes, and UFA. Table A2 classifies representative fallacies under all three. Pragma-dialectics and Toulmin's field-dependence, also discussed in §6.6, are omitted from the tables for readability; their relationship to the framework is procedural rather than taxonomic, and is treated in the prose. Map entry numbers (#n) refer to the companion Universal Fallacy Map (Second Edition, https://doi.org/10.5281/zenodo.21880635); R-numbers and §-references in the UFA column refer to the AOML v2.2 specification.

Table A1. Organizing principles of the three traditions.

Dimension	Aristotelian / traditional formal–informal	Walton's argumentation schemes	AOML / UFA (v2.2)
Primary organizing principle	Form of the argument (validity) plus named surface errors	Presumptive reasoning patterns plus critical questions	Boundary violations of coherent distinction-making (five types)
Source of the taxonomy	Catalogue from observed errors (Sophistical Refutations onward)	Catalogue of common schemes plus defeasibility conditions	Derived from the AOML constraint geometry (axes, operators, methods, levels, the FRAME) and the CLCA evidence
Number of primitive categories	Typically two (formal / informal) plus a long list of named fallacies	Dozens of schemes; no fixed small set of primitives	Exactly five types (plus compounds); closed under the geometry
Treatment of compounds	Usually forced into one named fallacy	Often treated as scheme mixtures or separate schemes	First-class: analyzed as co-occurring violations (e.g., Type 1+5,

			Type 2+4)
Meta-level unification	Limited (e.g., relevance fallacies grouped loosely)	Limited; each scheme largely independent	Strong: the single schema M[A] ⇸ A generates a large family (R5/R6/R8/R9)
Predictive / generative power	Low: new fallacies added ad hoc	Moderate: new schemes can be proposed	High: the space of structural error is calculable from the geometry (§3); any new structural error must reduce to one or more of the five types
Link to success conditions	Validity only (form-level); no axis-relative success conditions	Defeasibility conditions per scheme; no global dual with truth conditions	Dual with axis-relative truth (Paper IV)
Cross-substrate applicability	Natural-language argument	Natural-language argument, with computational implementations of schemes; no claim about non-argumentative systems	Explicitly claimed for any distinction-making system, human or engineered (§4; Paper III)
Falsifiability of the inventory	Open-ended list	Open-ended list of schemes	Closed: a sixth primitive type requires a fifth condition for coherent distinction-making (§3.3; Paper III)

Table A2. Representative fallacies under the three traditions.

Fallacy	Traditional / Aristotelian label	Walton treatment (typical)	AOML / UFA classification	Structural advantage of UFA

Ad Hominem	Informal (relevance)	Ad hominem scheme plus critical questions	Type 1: SINC → VER (map #1)	Axis transfer made explicit
Post Hoc Ergo Propter Hoc	Informal (causal)	Causal argumentation scheme	Type 2: CAUSE misapplied (map #25)	Operator licensing violation
Appeal to Authority	Informal (relevance)	Argument from expert opinion scheme	Type 3: habitat mismatch (map #48) or Type 2: EVID[A] ↛ A, R9 (map #100)	Distinguishes two structurally distinct errors under one traditional name
Hasty Generalization	Informal (induction)	Argument from example / generalization	Type 2+4: inductive license root, INSTANCE → CATEGORY locus (map #66)	Root/locus separation made explicit
False Dilemma	Informal (presumption)	Argument from alternatives / black-and-white	Type 5: BINARY imposed (map #77)	Opposition-topology precondition
Straw Man	Informal (relevance)	Straw man scheme	Type 5: OPPOSITION distorted (map #79)	Frame mis-specification
Appeal to Probability	Informal, or formal (modal)	Often under probabilistic schemes	Type 2: DEG[VER] ↛ VER, R5 (map #34)	Meta-axis substitution family
Sunk Cost	Rarely classical	Practical reasoning / sunk-cost scheme	Type 2+4: TEMP substitution plus PAST → FUTURE, R6 (map #65)	Compound made transparent
Forgery / Truthiness /	Rarely treated as fallacies of	Not standard	Type 2: LIKE[A] ↛ A,	The deception forms unified

Hypocrisy	argument	schemes	R8 (spec §V; inferential form at map #72)	with the fallacies under one schema
Circular Reasoning	Formal / begging the question	Begging the question scheme	Type 2: identity treated as derivation (map #28)	Operator overextension
Whataboutism	Modern informal	Often under tu quoque or distraction	Type 1+5: SINC → VER plus CONTRAST (map #92)	Compound analysis explains the rhetorical force

The contrasts reduce to three. The classical tradition has the long empirical record and the pedagogical clarity, but remains largely a surface catalogue: the formal/informal cut is a useful first pass, yet it does not explain why the informal errors cluster as they do, generate predictions, or handle compounds systematically. Walton's schemes bring pragmatic and dialectical sensitivity, and the critical questions make defeasibility explicit, but the inventory remains open-ended and scheme-by-scheme, with no derived claim that the space of structural failures is finite or closed. UFA's five types are derived from the conditions of coherent distinction-making (Paper III) rather than observed and catalogued; compounds are first-class; the meta-axis substitution family supplies the generative unification; the framework is dual with the success side, truth on each axis (Paper IV); and the architecture is projected into operational systems (§4), with engineering development registered as future work (§6.7). The current limitation is stated in §6.5: the underlying CLCA regularities are candidates pending native-speaker validation, and independent uptake of the framework is early; comparison tables of this kind are part of the demonstration work, not a substitute for it.