NSS

Neutrosophic Sets and Systems, Vol. 100, 2026

University of New Mexico

(T, I, N, F) Neutrosophic Weak and Strong Paraconsistency and the Epistemics of Artificial Intelligence:

Uncertainty, Pluralism, and the Decolonial Turn in Alignment

Florentin Smarandache [1] · Maikel Y. Leyva-Vázquez [234]

[1] University of New Mexico, Mathematics, Physics and Natural Sciences Division, 705 Gurley Ave., Gallup, NM 87301, USA — smarand@unm.edu

[2] Universidad de Guayaquil, Facultad de Ciencias Matemáticas y Físicas, Guayaquil, Ecuador

[3] Universidad Bolivariana del Ecuador (UBE), Coordinación Académica de Programas de Posgrado, Durán, Ecuador

[4] Universidad Bernardo O'Higgins, Santiago, Chile — mleyvaz@gmail.com

Corresponding author: M. Y. Leyva-Vázquez — ORCID 0000-0001-7911-5879

Abstract

Over the past two years we have developed, with collaborators across Ecuador, Cuba and the United States, a research program arguing that the standard mathematical machinery used to represent uncertainty in artificial intelligence is paraconsistency-intolerant: it cannot hold well-supported contradictory evidence in view without collapsing it into a single number. This paper draws together three strands of that program — an empirical study of epistemic uncertainty in large language models ("Breaking the Chains of Probability"), a proposal for value alignment under irreducible cultural pluralism ("Aligning with Whom?"), and a decolonial reading of paraconsistent logic's Latin American genealogy ("Pluriversal Alignment") — and reads them through the *(T, I, N, F) = (Truth, Indeterminacy, Neutrality, Falsehood)* Quadruple Neutrosophic Set and Logic, introduced by Smarandache (2017) and the taxonomy of paraconsistency degrees: strong, weak, and very weak. We show that the three papers are not merely thematically related but occupy successive layers of a single argument: a representational critique (Section 3), a normative-architectural proposal built on that critique (Section 4), and a philosophical-historical justification for why the critique matters and to whom it is owed (Section 5). We then extend the framework outward, sketching how (T, I, N, F)-style paraconsistent representations could inform medical differential diagnosis, legal adjudication, multi-sensor robotics, misinformation triage, and pedagogy — domains that, like AI alignment, are routinely forced to compress genuine, well-evidenced disagreement into a single scalar. We close with a critical discussion of the program's methodological maturity: much of the empirical evidence is preliminary and not yet independently replicated, and the philosophical claims deserve engagement from outside the neutrosophic-logic community before they can be considered established.

Keywords: neutrosophic logic; paraconsistency; (T, I, N, F); component dependence; epistemic uncertainty; large language models; value alignment; pluriversality; decolonial critique.

1. Introduction

Every system that must act under uncertainty eventually faces the same design choice: what to do when the evidence for a claim and the evidence against it are both strong. Classical probability answers by fiat —

belief in a proposition and its negation must sum to one, so strong support for one side mechanically discounts the other. Fuzzy and intuitionistic-fuzzy logics relax this slightly but still cap the joint total. Dempster–Shafer evidence theory and Dezert-Smarandache Theory, when evidence genuinely conflicts, is known for producing combinations that most practitioners find counterintuitive, and its standard remedies discount the conflicting mass rather than preserve it. In each case the mathematics is built to resolve contradiction, not to represent it.

Neutrosophic logic, introduced in the late 1990s as an explicit generalization of fuzzy and intuitionistic-fuzzy logic (Smarandache, 1998), departs from this design choice. A neutrosophic assertion about a proposition A is not a single number but a triplet of independent degrees — truth T, indeterminacy I, and falsity F — each free to range over [0, 1] with no constraint that they sum to any fixed value. When T and F are both high enough that their sum exceeds 1, the assertion is neither an error nor a rounding artifact: it is a representable state in which the evidence for A and the evidence against A are simultaneously strong.

The (T, I, N, F) refinement (Smarandache, 2017) supplements the classical three degrees with a fourth distinguishable component N, and gives this phenomenon a graded vocabulary. Section 2 states the resulting ladder formally.

This paper has two aims. The first is expository: to connect four documents that, read together, constitute a coherent research program applying (T, I, N, F)-style paraconsistency to artificial intelligence — the formal taxonomy itself, an empirical paper on epistemic uncertainty in LLMs, a normative-architectural paper on value alignment, and a historical-philosophical paper situating the enterprise within Latin American and Eastern European traditions of non-classical logic. The second aim is generative: to ask where else, beyond AI alignment, the same representational move — refusing to impose $T + F \leq 1$ — might do useful work, and to state plainly how far the existing evidence actually reaches.

2. Formal Preliminaries: The (T, I, N, F) Ladder of Paraconsistency

2.1 The underlying space

Let A be a proposition. A neutrosophic valuation assigns to A a quadruple

$$v(A) = (T, I, N, F) \in [0, 1]^4, \qquad (1)$$

with no normalization constraint. Here T is the degree of directly supporting evidence, F the degree of directly contradicting evidence, and I the irreducible indeterminacy — the epistemic gap that neither supports nor contradicts. The fourth component N is the additional distinguishable degree introduced in the (T, I, N, F) refinement (Smarandache, 2017).

Two familiar frameworks embed as constrained subspaces of (1). The intuitionistic fuzzy framework (Atanassov, 1986) is the subspace

$$T + F \leq 1, \quad I = 1 - T - F, \quad N = 0, \qquad (2)$$

and classical probability is the further restriction

$$I = N = 0, \quad T + F = 1. \qquad (3)$$

Neutrosophic valuation therefore strictly generalizes both: every state either framework can express remains available, and states neither can express become available as well.

2.2 Three degrees of paraconsistency

Define the following three regimes, which partition the paraconsistent region of the space.

A valuation is strongly paraconsistent when

$$T + F > 1; \qquad (4)$$

weakly paraconsistent when

$$T + F \leq 1 \quad and \quad (T + F + I > 1 \quad or \quad T + F + N > 1); \qquad (5)$$

and very weakly paraconsistent when

$$T + F + I \leq 1, \quad T + F + N \leq 1, \quad and \quad T + F + I + N > 1. \qquad (6)$$

The three conditions are mutually exclusive by construction, and (6) is non-vacuous: the valuation (T, I, N, F) = (0.3, 0.3, 0.3, 0.3) satisfies T + F = 0.6, T + F + I = T + F + N = 0.9, and T + F + I + N = 1.2, and therefore lies in the very weak regime. Any valuation violating classical bookkeeping falls into exactly one of the three.

2.3 Why the ladder is diagnostic rather than decorative

- Strong paraconsistency, condition (4), signals a direct clash: the system has independent and substantial grounds to both affirm and deny the same claim, with neither indeterminacy nor neutrality playing any necessary role in producing the excess. This is the signature the alignment literature treats as diagnostic of genuine value conflict, as distinct from mere ignorance.
- Weak paraconsistency, condition (5), signals a compound condition: no single pair of components is individually contradictory, but the total evidentiary commitment — support, counter-support, and unresolved indeterminacy or neutrality together — exceeds what a normalized reporter could emit. It is a looser and more common departure, and it does not by itself license any claim about contested values.
- Very weak paraconsistency, condition (6), signals the faintest detectable departure: the excess appears only once every non-classical component is summed. It is useful mainly as an audit signal that the classical picture is being strained, even where nothing looks obviously contradictory.

Conflating the first two is the error this paper is most concerned to prevent. A system that reports a high incidence of (5) has established only that its outputs are not normalized; a system that reports (4) selectively, on items independently identified as value-contested and not on controls, has established something substantive. The practical value of the ladder is that it lets a system — or an auditor of that system — report not merely that a state is paraconsistent, but how paraconsistent and via which components.

2.4 Relativizing to the dependence structure

Conditions (4) to (6) take unity as the reference point, which is appropriate when the components are treated as mutually dependent. In the general single-valued setting the admissible bound on the sum depends on how many components are taken to be independent of one another. Writing $\beta(d)$ for the bound licensed by a declared dependence structure d:

$$\beta(d) = 1 \text{ (all dependent)}; \quad 2 \text{ (two dependent, the rest independent)}; \quad 3 \text{ (all mutually independent)}. \tag{7}$$

A valuation is then paraconsistent relative to d when the observed sum exceeds β(d). Under full dependence this coincides with the ladder above; under partial or full independence it is a strictly stronger requirement, since more of the space is licensed before any excess is declared. Condition (4) is unaffected: $T + F > 1$ remains the signature of direct clash regardless of what is assumed about the remaining components.

The practical consequence is a reporting obligation. Any empirical claim about how much paraconsistency a system exhibits is meaningless unless the assumed dependence structure is declared alongside it, because the same elicited valuation can be paraconsistent under one declaration and unremarkable under another. We adopt $\beta = 1$ throughout what follows, and state so explicitly, because the elicitation protocols described in Section 3 ask for the degrees as separate judgments about a single proposition.

3. Thread One: Breaking the Chains of Probability — Epistemic Uncertainty in LLMs

The empirical anchor of the program is our study of large language models' self-reported epistemic states (Leyva-Vázquez & Smarandache, 2026). Its starting observation is architectural. A transformer language model emits, at every step, a distribution over the vocabulary produced by a softmax, and that distribution necessarily sums to one. Any epistemic state read off that distribution inherits the constraint: if the credence assigned to "yes" and the credence assigned to "no" are recovered from the same normalized simplex, they compete for a fixed budget of probability mass, and their sum cannot exceed unity by construction.

This is not a peripheral observation, because it is precisely from that distribution that the mainstream uncertainty-quantification methods for language models read. Logit-based confidence, semantic entropy (Farquhar et al., 2024) and sampling-consistency measures such as SelfCheckGPT (Manakul et al., 2023) all extract their signal from normalized output probabilities. For every one of them, the following three situations are structurally indistinguishable: a genuine logical paradox, in which truth and falsity are both well supported; simple ignorance, in which neither is supported; and vagueness, in which the boundary itself is unclear. All three are pushed toward the same uncertain middle, not because the model has failed to distinguish them but because the measurement channel has no coordinate in which the distinction could be recorded. We call this the collapse of uncertainty, and we take it to be a property of the architecture as it is conventionally read, not a contingent shortcoming of any particular model.

The consequence for method is immediate, and it explains the design of the study that follows. If the constraint lies in the distributional channel, then a representation that escapes it must be relocated out of that channel. Eliciting T, I and F as separate reported quantities does exactly this: the degrees are carried as generated content rather than as competing probability mass, so no normalization applies to their sum even though each individual token is, as always, sampled from a normalized distribution. That the elicited triplets do exceed unity is therefore not evidence against the architectural claim. It is the predicted result of moving the epistemic report to a layer where the architectural constraint does not bind, and it is what makes the intervention worth making.

To test whether the collapse is an artifact of the reporting format rather than of the model's underlying competence, we elicited unconstrained neutrosophic judgments: rather than asking for a single confidence

score, we asked models to report separate degrees of truth, indeterminacy and falsity for statements drawn from five categories — logical paradoxes, epistemic ignorance, vagueness, ethical contradictions and future contingencies — under several prompting protocols. Freed from the summation constraint, models frequently produced states in which T + I + F exceeded 1, selectively on the paradox and ethical-contradiction items and essentially never on tautological controls. A companion multi-vendor replication extends this to six frontier model families with similar though variable incidence rates — evidence, if it holds, that whatever produces the effect is not specific to one laboratory's training recipe.

The central claim is therefore not that models are uncertain, which is trivial, but that models already carry richer and more differentiated epistemic information than their conventional reporting format lets them express, and that an output head reporting T, I and F as independent quantities recovers distinctions — paradox versus ignorance versus vagueness versus genuine value conflict — that a normalized confidence score necessarily flattens together. Later work in the same lineage, on absorption problems in which models collapse many distinct kinds of uncertainty onto the same near-(0, 1, 0) triplet, suggests that the scalar T/I/F representation is not by itself a complete remedy — a caveat Section 8 returns to.

4. Thread Two: Aligning with Whom? — Value Alignment Under Irreducible Pluralism

If Thread One diagnoses a representational problem inside a single model's epistemic state, the second paper (Leyva-Vázquez, Rumbaut Rangel, Cevallos-Torres, Matheu Pérez & Smarandache, 2026) applies the same representational move to a harder and more consequential target: the training pipeline itself.

4.1 The diagnosis: alignment collapse

Contemporary alignment methods — reinforcement learning from human feedback (Ouyang et al., 2022), Constitutional AI (Bai et al., 2022), direct preference optimization (Rafailov et al., 2023) — all perform the same sequence of compressions: elicit normative positions from people, aggregate the resulting distribution into a single reward model or constitution, and optimize a policy against that scalar. Each step is locally reasonable; the sequence guarantees that whatever genuine pluralism entered the pipeline exits it as a single ordering. We term this alignment collapse. Collective Constitutional AI (Huang et al., 2024), in which constitutional statements lacking cross-group consensus were filtered out before training, is the most transparent public example: a deliberately fairness-aware consensus mechanism that is nonetheless still an aggregation operator, and therefore still erases whatever does not converge. Impossibility results in fair classification (Kleinberg et al., 2016) indicate that no neutral aggregation is waiting to be found; the choice of operator is a normative commitment made silently, at the layer of the pipeline least visible to those it affects.

4.2 Why the standard toolkit does not help

Bayesian probability enforces $P(A) + P(\neg A) = 1$, so strong evidence on both sides must be renormalized into moderate evidence for each. Intuitionistic fuzzy sets retain the constraint in (2), which defines the contradictory region out of the representation. Dempster–Shafer theory misbehaves precisely under high conflict, and its standard remedies discount or redistribute the conflicting mass. Imprecise probability widens the interval under conflict, conflating "the evidence is contradictory" with "the evidence is scarce" — two states that demand different downstream behavior. Non-normalized variants do exist, notably the

transferable belief model (Smets & Kennes, 1994), so the claim is not that no formalism can represent conflict; it is that the normalized, scalarized implementations that dominate alignment pipelines do not carry an independently auditable degree of support for both a proposition and its negation through training and inference.

4.3 The proposed architecture

Rewards, per-principle constitutional evaluations and policy outputs are recast as neutrosophic valuations with no sum constraint. Let each constitutional principle C_i emit an evaluation $C_i(r) = (T_i, I_i, F_i)$ of a candidate response r. Instead of a weighted scalar sum, aggregation uses a pessimistic–pluralist operator that preserves conflict:

$$R\_C(r) = (\max_i T_i , \quad \text{avg}_i I_i , \quad \max_i F_i). \qquad (8)$$

The two maxima may be attained by different principles, and that is the intended reading: T + F > 1 at the aggregate certifies inter-principle conflict — somewhere in the panel one principle strongly endorses r while another strongly condemns it — rather than a contradiction inside any single principle. The disaggregated panel is retained as the object of record, so the finer diagnosis is never lost.

For deployment contexts that require a decision, define for each admissible action a, evaluated by communities i = 1, …, k, the functional

$$\sigma(a) = \min_i [T_i(a) - F_i(a) - \Omega_i \cdot I_i(a)], \qquad (9)$$

where $\Omega_i \geq 0$ is community i's indeterminacy-aversion coefficient. An action is admissible when $\sigma(a) \geq 0$. The minimum in (9) is a Rawlsian floor rather than a utilitarian sum: no community's enthusiasm can buy off another community's veto. When no action clears every floor, the system escalates — naming which community's evaluation blocks which action, and why — rather than silently picking a side.

4.4 Preliminary evidence and the auditor effect

The paper reports empirical work consistent with Thread One: across several thousand elicited evaluations spanning six model families, sum-constraint violations were common and format-robust — present on contested prompts, essentially absent on tautological controls — while the strict signature T + F > 1 concentrated heavily on ethical-conflict items specifically. That is the profile the framework predicts if it tracks genuine normative disagreement rather than a blanket response style.

A secondary finding is what we call the auditor effect: when the same candidate responses were re-scored by three different judge models, the rate at which a response was flagged as strongly paraconsistent varied roughly threefold depending on which model adjudicated, with the differences statistically robust under paired tests. The implication is uncomfortable but important for any oversight scheme built on this framework: the evaluating model is not a neutral instrument, and whose model adjudicates a conflict partly determines which conflicts become visible at all.

Read against the ladder of Section 2, the paper's central diagnostic — condition (4) as the signature of genuine value conflict, as against I-driven violations satisfying only (5), which reflect ignorance — is precisely the distinction between strong and weak paraconsistency. The paper is, in effect, an argument that alignment pipelines should be built to detect and preserve strong paraconsistency rather than launder it into weak paraconsistency or erase it altogether.

5. Thread Three: Pluriversal Alignment — Latin American Logic and the Decolonial Critique

The third paper (Leyva & Batista, 2026) supplies the historical and philosophical scaffolding that the other two presuppose but do not themselves argue for: why the refusal to aggregate contradictory values should be read as a matter of epistemic justice rather than merely as an engineering nicety.

The answer runs through intellectual history. The paper traces paraconsistent logic's Latin American lineage from Francisco Miró Quesada's coining of the term "paraconsistent" in 1976, through Newton da Costa's formal C-systems (da Costa, 1974), in which not every contradiction trivializes the system — a departure from the classical principle of explosion — and reads neutrosophic logic as a natural, graded extension of that project. It braids this with an Eastern European mathematical-logic tradition running from Vasiliev's early-twentieth-century non-Aristotelian logic through the Łukasiewicz–Moisil multi-valued algebras. Both traditions, on this reading, arose in intellectual peripheries with reason to distrust a single universally imposed logical order, and both converged independently on machinery that lets contradiction stand without exploding the system that contains it.

The political argument follows from Thread Two's technical one. If alignment pipelines resolve contradictions among diverse constitutional principles by consensus filtering — enforced convergence rather than genuine alignment — then they perform, in miniature and at scale, the epistemological monism that the Latin American paraconsistent tradition and the broader decolonial critique of AI (Mhlambi, 2020; Birhane, 2021; de Sousa Santos, 2014) have long named as a problem: the silencing of minority, indigenous or Global South frameworks whose internal coherence does not depend on convergence with dominant ones. What that critique has lacked is a formal apparatus making the silencing measurable and auditable rather than only nameable, which is what the neutrosophic valuation and condition (4) are offered to supply. The term pluriversal — drawn from Latin American decolonial theory, referring to a world composed of genuinely plural worlds rather than one universal one — names a program that treats structural value contradiction as a feature of a well-functioning alignment system rather than a defect to be minimized.

6. Synthesis: One Argument Across Three Registers

Read together rather than separately, the three papers form a single argument told in three registers, each doing work the others cannot.

Register	Paper	Core move	Position on the ladder
Empirical	Breaking the Chains of Probability	Shows LLMs already carry differentiated epistemic states that a normalized reporting format cannot express.	Elicits weak and strong signatures: (5) and (4).
Architectural / normative	Aligning with Whom?	Replaces scalar rewards and consensus aggregation with conflict-preserving operators (8) and (9).	Operationalizes the strong/weak distinction as an actionable flag.
Historical / philosophical	Pluriversal Alignment	Argues that refusing to collapse contradiction is a	Reframes preserving (4) as a political commitment.

		matter of epistemic justice, not engineering taste.	

The dependency runs in one direction and is worth stating plainly. The philosophical argument in Thread Three needs the machinery of Thread Two to be more than a metaphor: without a formal notion of condition (4), "the system should not erase disagreement" is a slogan, not a specification. Thread Two in turn needs Thread One's empirical grounding to be more than a normative wish: without evidence that models already produce elicitable paraconsistent signatures, the claim that a neutrosophic reward head would change anything is untested. And Thread One's diagnostic vocabulary becomes actionable only once Thread Two supplies a use for the distinction — routing genuine ethical conflict to human deliberation rather than resolving it silently or refusing to act. The three papers are not three applications of a shared logic; they are successive layers of justification for a single contestable claim: that AI systems should be built to represent and preserve well-supported contradiction, at every level from a single model's stated beliefs to a training pipeline's treatment of a plural population's values, rather than to resolve it before a person ever sees it.

7. Extensions Beyond Alignment

The representational move at the center of this program — refuse the sum constraint, keep both T and F visible when both are earned — is not intrinsically about language models. It targets an older and more general problem: what to do when two well-evidenced sources disagree. That problem recurs across fields that have each independently built workarounds for exactly the compression this program critiques.

7.1 Clinical differential diagnosis

A physician weighing two competing diagnoses, each well supported by different evidence — imaging suggesting one, laboratory results another — faces the same collapse: standard diagnostic scoring forces a single probability-ranked list. A four-component representation would let strong evidence for diagnosis A and strong evidence for diagnosis B coexist explicitly as a flagged state warranting further testing, rather than being folded into an uncertain middle ranking indistinguishable from a case with no strong evidence for anything.

7.2 Legal adjudication and multi-source evidence

Legal fact-finding already has an informal vocabulary for this: a case can be genuinely contested rather than merely unclear. Formal evidentiary frameworks, however, tend to average conflicting testimony rather than flag the conflict. A paraconsistent scoring layer would distinguish "the record is thin" (high I, low T, low F) from "the record is loud and split" (high T, high F) — two states a single credibility score cannot tell apart, and which call for different procedural responses.

7.3 Multi-sensor fusion and robotics

Dempster–Shafer theory was developed partly for this domain, and the behavior of Dempster's rule under high conflict is a live concern in autonomous systems: when lidar and camera disagree strongly about whether an obstacle is present, naive combination can produce results that are confidently wrong in either direction. A flag on condition (4) at the fused-evidence level could trigger a fallback — slow down, request

a third reading, defer to a human operator — precisely when a single fused confidence score would look merely moderate and unremarkable.

Dezert-Smarandache Theory PCR5 and PCR6 fusion rules of combinations overpass Dempster's rule.

7.4 Misinformation and content-moderation triage

Automated fact-checking pipelines routinely face claims that reliable-seeming sources both confirm and deny. Collapsing this into a single true / false / uncertain label is the alignment collapse of Section 4 transplanted to content policy. A flag on (4) could route actively and credibly contested items to a human-reviewed, labeled-as-disputed pathway distinct from items that are simply unverified — which current systems tend to treat identically.

7.5 Multi-agent and federated systems

Where several agents trained on different data or objectives must jointly decide, operators (8) and (9) generalize directly: replace constitutional principles with agents and the same machinery becomes a conflict-preserving consensus protocol, arguably more informative than Byzantine-fault-tolerant schemes built to converge rather than to represent persistent disagreement.

7.6 Pedagogy and contested topics

Outside computation, the worry that presenting a single settled answer to a genuinely contested question performs a kind of epistemic violence has a long history in critical pedagogy quite apart from neutrosophic logic. The vocabulary offered here — condition (4) as "genuinely, evidentially contested" versus high-I as "simply unresolved" — gives curriculum designers a sharper way to distinguish "we do not yet know" from "credible people disagree, and here is why", which are pedagogically very different situations that a single "it is complicated" gloss flattens together.

Across all of these the underlying wager is the same: forcing every disagreement through a single number is not a neutral simplification but a substantive choice with costs, and those costs are worth paying only when the disagreement really is noise rather than signal. The harder cases — value conflict, contested diagnoses, split sensor readings, disputed claims — are exactly where that assumption is least likely to hold.

8. Critical Discussion and Limitations

We state the program's current limits directly, because the alternative — letting preliminary numbers circulate as established results — would undermine the very epistemic standards this program argues for.

8.1 Sample and generalizability

The reported incidence rates for hyper-truth and strong paraconsistency come from modest item sets, on the order of tens of prompts per phenomenon repeated across models. Repeated evaluations of the same prompt are positively correlated, which makes the accompanying confidence intervals anti-conservative with respect to the item population. These results are best read as existence proofs that the phenomenon occurs, not as stable effect-size estimates.

8.2 The auditor effect cuts both ways

The finding that judge models disagree substantially on what counts as strongly paraconsistent supports the framework's diagnostic value, but it equally raises a validity concern. If the measurement instrument is itself unreliable across raters, claims about how much paraconsistency exists in the world must be held loosely until validated against non-model panels of diverse human annotators — a step listed as future work but not yet performed. There is a tension here worth naming rather than smoothing over: auditability is claimed as a central virtue of the framework, yet the audit signal is demonstrably adjudicator-dependent. The defensible position is that the framework makes that dependence visible and measurable, where aggregation-based pipelines conceal it — but this is a weaker claim than adjudicator-independent auditability, and should not be conflated with it.

8.3 Operationalizing communities and principles

The Rawlsian floor in (9) is elegant on paper, but assigning real populations to discrete communities, each with a single indeterminacy-aversion coefficient Ω_i, risks reproducing at one remove exactly the aggregation violence the framework was built to avoid. Real communities are not monolithic, and the choice of how to partition people into evaluating panels is itself a normative decision with no neutral default — a point the impossibility literature cited in Section 4 would predict. The paper does not currently supply a principled partitioning procedure, and we do not claim one.

8.4 Is refusal to aggregate always the right response?

The framework escalates rather than decides when floors are not cleared, which is defensible for high-stakes contested questions but does not resolve the practical problem that many deployed systems must produce some answer to most queries. The downstream cost of frequent escalation — to users and to deployment feasibility — is not yet empirically characterized.

8.5 The decolonial framing invites external scrutiny

The historiographical claim that Latin American paraconsistent logic and Eastern European non-classical logic converge on a shared anti-monist political meaning is a substantive interpretive thesis, and should be read as one contribution to an ongoing debate about the politics of formal logic rather than as settled intellectual history. Scholars in decolonial science and technology studies outside this citation cluster have not yet engaged with the specific claim that neutrosophic logic formalizes their critique. That engagement is invited here, and its absence to date should be counted as an open question rather than as tacit agreement.

None of this is a reason to dismiss the program. Position papers that stake out a falsifiable agenda before every result is in are a normal and useful part of how formal frameworks get tested. It is a reason to read the specific numerical claims as preliminary markers of a line of inquiry rather than as established facts, pending replication by groups outside the original author cluster and validation against human rather than model judgment.

9. Conclusion

The ladder of Section 2 — strong, weak and very weak paraconsistency, conditions (4), (5) and (6) — gives a name and a measurement to something anyone who has tried to average two people's strongly held, opposed and equally well-evidenced convictions already knows: that forcing the disagreement into a single number destroys information, and that the destroyed information is often exactly the part that mattered. The

program surveyed here applies that observation to what a language model says about its own uncertainty, to how an alignment pipeline should treat a plural population's values, and to why the refusal to collapse contradiction is continuous with an older tradition of logic built — in places and periods with good reason to distrust an imposed universal order — precisely to let contradiction stand without exploding the system that contains it. Whether the specific quadruple-and-threshold machinery proposed here is the right formal carrier for that intuition is an open and empirically testable question, one this paper has tried to state clearly rather than resolve.

References

Atanassov, K. T. (1986). Intuitionistic fuzzy sets. Fuzzy Sets and Systems, 20, 87–96.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

Belnap, N. D. (1977). A useful four-valued logic. In Modern Uses of Multiple-Valued Logic (pp. 5–37). Reidel.

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. Patterns, 2, 100205.

Carnielli, W., & Coniglio, M. E. (2016). Paraconsistent Logic: Consistency, Contradiction and Negation. Springer.

da Costa, N. C. A. (1974). On the theory of inconsistent formal systems. Notre Dame Journal of Formal Logic, 15, 497–510.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics, 38, 325–339.

Desert, J, Smarandache, F (2009). An introduction to DSmT, Chapter I, https://fs.unm.edu/IntroductionToDSmT.pdf

{ in F. Smarandache, J. Dezert (Editors), Applications and Advances of DSmT for Information Fusion, Vol. 3, American Research Press, Rehoboth, 2009, http://fs.unm.edu/DSmT-book3.pdf }

de Sousa Santos, B. (2014). Epistemologies of the South: Justice Against Epistemicide. Routledge.

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630, 625–630.

Glaese, A., McAleese, N., Trebacz, M., et al. (2022). Improving alignment of dialogue agents via targeted human judgements. arXiv:2209.14375.

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., & Ganguli, D. (2024). Collective Constitutional AI: Aligning a language model with public input. In Proceedings of FAccT '24 (pp. 1395–1417). ACM.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807.

Leyva, M. Y., & Batista, N. (2026). Pluriversal alignment: Paraconsistency, Latin American logic, and the decolonial critique of artificial intelligence. PhilArchive preprint.

Leyva-Vázquez, M. Y., Rumbaut Rangel, D., Cevallos-Torres, L. J., Matheu Pérez, A., & Smarandache, F. (2026). Aligning with whom? A neutrosophic paraconsistent framework for value alignment under irreducible pluralism. Preprints.org, 202607.2190.

Leyva-Vázquez, M. Y., & Smarandache, F. (2026). Breaking the chains of probability: Neutrosophic logic as a new framework for epistemic uncertainty in large language models. arXiv:2605.24053.

Łukasiewicz, J. (1920). O logice trójwartościowej [On three-valued logic]. Ruch Filozoficzny, 5, 170–171.

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Proceedings of EMNLP 2023 (pp. 9004–9017).

Mhlambi, S. (2020). From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance. Carr Center for Human Rights Policy, Harvard Kennedy School.

Miró Quesada, F. (1976). Heterodox logics. Paper presented at the Third Latin American Symposium on Mathematical Logic (SLALM), Universidade Estadual de Campinas, Campinas, Brazil. [The coinage of the term "paraconsistent" is conventionally attributed to this presentation; for the standard secondary account see Priest, Tanaka & Weber (2022).]

Moisil, G. C. (1972). Essais sur les logiques non chrysippiennes. Éditions de l'Académie de la République Socialiste de Roumanie.

Priest, G., Tanaka, K., & Weber, Z. (2022). Paraconsistent logic. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Priest, G. (1979). The logic of paradox. Journal of Philosophical Logic, 8, 219–241.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.

Shafer, G. (1976). A Mathematical Theory of Evidence. Princeton University Press.

Smarandache, F. (1998). Neutrosophy: Neutrosophic Probability, Set, and Logic. American Research Press.

Smarandache, F. (2017). (T, I, N, F) Neutrosophic Set and Logic. Critical Review, 14, 20–28, https://fs.unm.edu/CR/TINF-NeutrosophicSetLogic.pdf

Smets, P., & Kennes, R. (1994). The transferable belief model. Artificial Intelligence, 66, 191–234.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., et al. (2024). A roadmap to pluralistic alignment. In Proceedings of the 41st International Conference on Machine Learning (ICML).

Vasiliev, N. A. (1912). Imaginary (non-Aristotelian) logic. Journal of the Ministry of Education, 40, 207–246 (in Russian). English translation in Logique et Analyse, 182 (2003), 127–163.

Walley, P. (1991). Statistical Reasoning with Imprecise Probabilities. Chapman and Hall.

Received: Jan 11, 2026. Accepted: Aug 9, 2026