Neurotransparency: An Epistemic Primitive for AI–Human Collaboration

Author: Shawn C. Wright
Affiliation: Waveframe Labs / Aurora Research Initiative
DOI: [10.5281/zenodo.17013612]

Abstract

Neurotransparency is introduced as a methodological requirement ensuring that every claim-affecting inference—whether generated by human or synthetic cognition—is explicitly attributable, traceable, and reconstructible.
It establishes the epistemic boundary condition for reproducible AI-assisted research: **no reasoning may influence a claim unless its origin and evidence path are recorded.**

As an epistemic primitive, neurotransparency replaces interpretive trust with deterministic auditability.
It underlies both the **Aurora Workflow Orchestration (AWO)** methodology and its enforcement layer **CRI-CORE**, forming the ethical and procedural foundation for transparent AI-human reasoning systems.

---

1. Definition

> **Neurotransparency (n.)** — the principle that every inference influencing a claim must be attributable to a declared role and recorded in a durable artifact such that the reasoning sequence can be deterministically reconstructed.

Formally, this constitutes a **traceability invariant**: [R_{claim} = f(Evidence, Role, Hash)] where each reasoning contribution is recorded as a tuple *(evidence pointer, role identifier, content hash)* and becomes a permanent element of the project's provenance ledger.

---

2. Motivation

Modern AI-assisted research collapses traditional authorship boundaries: reasoning steps occur across humans, language models, and automated validators.
Without a deterministic attribution framework, knowledge becomes non-verifiable once the underlying model or context changes.

Existing reproducibility standards focus on data and code.
Neurotransparency extends reproducibility to cognition itself—capturing not just *what* was done but *why* a conclusion was reached, and by whom (or what).

This shift transforms epistemic credibility from a social construct to an evidentiary one.

---

3. Implementation in AWO

AWO enforces neurotransparency through three structural guarantees:

Mechanism	Artifact	Enforcement
Role Attribution	`/logs/workflow/`, `/decisions/ADR-NNNN`	Every inference recorded with declared role (Orchestrator, Auditor, Synthesizer, etc.)
Hash Continuity	`SHA256SUMS.txt`	Every reasoning artifact cryptographically linked to its origin context
Attestation Linkage	`approval.json`	All validated claims must reference their evidence pointers and hash records

Failure to provide evidence linkage or role attribution constitutes a **non-conformance** event under §1.6 of the AWO Method Specification.

---

4. Enforcement Pathway (CRI-CORE Context)

The upcoming **CRI-CORE** enforcement layer operationalizes neurotransparency through executable schemas:

Module	Function	Schema Reference
`neurotransparency.schema.json`	Defines minimal fields: role, evidence path, hash, timestamp	AWO `/schemas/`
`reasoning_ledger.py`	Serializes reasoning steps and assigns deterministic IDs	CRI runtime
`attestation_validator.py`	Verifies that every claim references a valid reasoning record	CRI enforcement gate

CRI-CORE treats neurotransparency as a *first-class constraint*: builds fail if any claim-affecting artifact lacks verifiable provenance.

---

5. Ethical and Epistemic Implications

Principle	Classical Science	Neurotransparent Research
Authorship	Singular, human	Distributed, role-based

Principle	Classical Science	Neurotransparent Research
Reproducibility	Experimental and data-centric	Cognitive and reasoning-centric
Verification	Peer review	Deterministic audit
Failure Mode	Misinterpretation	Missing evidence pointer

By embedding reasoning transparency into the research substrate, AWO and CRI-CORE redefine *trust* as *trace*.
This enables a post-institutional form of epistemic governance where credibility is computed, not declared.

---

6. Use Cases

1. **Automated Literature Synthesis**
   Every summarized claim must cite its reasoning path and originating model role.
2. **Scientific Simulation Governance**
   Parameter selection and model calibration decisions logged as reasoning events, enabling full reconstruction.
3. **Policy or Ethics Review**
   Deliberative reasoning among AI agents traceable to hash-bound evidence pointers.
4. **Cross-Model Audit**
   Multiple models perform the same reasoning task; differences become explicit through neurotransparency logs.

---

7. Relationship to Adjacent Concepts

Concept	Relation	Distinction
Explainability (XAI)	Post-hoc interpretation	Neurotransparency is pre-registered reasoning capture
Accountability	Governance outcome	Neurotransparency is evidentiary infrastructure
Provenance	Historical chain	Neurotransparency is cognitive provenance—who reasoned, not just who edited
Neurosymbolic AI	Hybrid reasoning architecture	Neurotransparency governs epistemic trace, not computation mode

---

8. Governance and Citation

- This doctrine is governed under **ADR-0017**.
- Future enforcement handled by **CRI-CORE v0.1+** via `neurotransparency.schema.json`.
- The doctrine itself is a **citable artifact** and may be referenced as:

“‘bibtex @misc{wright_neurotransparency_doctrine_2025, author = {Wright, Shawn C.}, title = {Neurotransparency Doctrine: An Epistemic Primitive for AI–Human Collaboration}, year = {2025}, version = {1.0.0}, institution = {Waveframe Labs / Aurora Research Initiative}, license = {CC BY 4.0}, orcid = {0009-0006-6043-9295}, doi = {10.5281/zenodo.17013612} }