GWP: Global Workspace Protocol

Consciousness-Inspired AI Collaboration Through Recursive Integration and Contradiction Detection

Centram Research | March 28, 2026 | Centram Whitepaper Series

Global Workspace Protocol — applying Baars' Global Workspace Theory to distributed AI systems.

The Theoretical Foundation

In 1988, cognitive scientist Bernard Baars proposed the **Global Workspace Theory** (GWT) of consciousness. The theory posits that the human brain consists of many specialized, parallel processors — visual cortex, auditory cortex, language areas, motor planning regions — each operating independently on their own inputs. Consciousness arises when these processors compete for access to a shared "global workspace," a kind of cognitive bulletin board where winning signals are broadcast to all processors simultaneously.

The power of this architecture is not in any single processor's capability, but in the **integration**. When you see a red ball rolling toward a cliff edge, your visual system recognizes the object, your physics intuition predicts the trajectory, your language system can name what's happening, and your motor system prepares to catch it. These processes happen in parallel, but your conscious experience is a unified whole — a single narrative integrating all of these inputs.

GWP applies this principle to distributed AI. In Centram's architecture, DNA clusters are the specialized processors. Each cluster generates and verifies responses independently. GWP is the global workspace — the integration layer that collects cluster outputs, detects conflicts between them, and drives the system toward a coherent, unified response.

The Integration Cycle

GWP operates in a recursive cycle with five phases:

The cycle works as follows: First, the workspace collects outputs from all active clusters. Then it synthesizes them into a candidate response and actively searches for contradictions between the outputs. If no contradictions are found, the synthesis is finalized and delivered. If contradictions exist, the workspace generates deepening directives and sends them back to the relevant clusters, triggering a new round. This continues for up to five cycles. If contradictions persist after all cycles, the workspace applies graceful degradation — delivering the best available synthesis with an explicit acknowledgment of the unresolved disagreements.

Phase 1: Collect

GWP gathers the verified outputs from all active DNA clusters. Each output comes with metadata: the cluster's domain specialization, the number of generators and verifiers involved, the confidence level, and the verification margin (how far above the 2/3 threshold the approval was).

Phase 2: Synthesize

The workspace builds a candidate synthesis — an integrated response that combines the contributions of all clusters. This is not simple concatenation. The synthesis process identifies overlap, resolves redundancy, and constructs a unified narrative from the diverse inputs.

Phase 3: Detect Contradictions

This is the critical phase. The workspace actively searches for disagreements between cluster outputs. A system that only synthesizes without checking for contradictions would be no better than averaging — it would smooth over genuine disagreements rather than resolving them.

Phase 4: Issue Directives

When contradictions are found, the workspace generates **deepening directives** and sends them back to the relevant clusters. This triggers a new round of inference with additional context about what other clusters have said and where the disagreement lies.

Phase 5: Converge

When no contradictions remain (or contradictions have been reduced below a threshold), the synthesis is finalized and delivered as the response.

Four Types of Contradictions

Not all disagreements are the same. GWP classifies contradictions into four categories, each requiring different resolution strategies:

Factual Contradictions

Cluster A says "The population of Tokyo is 14 million." Cluster B says "The population of Tokyo is 37 million." These are direct factual conflicts where at least one cluster is wrong (or they're using different definitions — city proper vs. metropolitan area). Resolution: request clarification of scope from both clusters, cross-reference with additional clusters if available.

Logical Contradictions

Cluster A concludes "Therefore, the investment is low-risk." Cluster B, using the same premises, concludes "Therefore, the investment is high-risk." The reasoning chains lead to opposite conclusions. Resolution: the workspace identifies the specific inference step where the chains diverge and issues directives asking each cluster to justify that step.

Semantic Contradictions

Cluster A and Cluster B appear to agree but are using the same terms with different meanings. For example, "efficiency" in an engineering context (energy output / energy input) versus "efficiency" in an economic context (Pareto optimality). These are the most subtle contradictions. Resolution: the workspace requests explicit definitions of key terms from each cluster.

Ethical Contradictions

Cluster A recommends a course of action that Cluster B flags as ethically problematic. These are not factual disputes but value conflicts. Resolution: the workspace does not attempt to resolve the ethical disagreement itself. Instead, it presents both perspectives transparently, clearly labeling the ethical dimension of the disagreement for the user.

The detection process compares every pair of cluster outputs across all four dimensions. For each pair, it checks factual consistency (do the claims align?), logical consistency (do the reasoning chains converge?), semantic alignment (are key terms used with the same meanings?), and ethical consistency (are there value conflicts?). Each detected contradiction is tagged with its type, the clusters involved, and a severity level that determines whether it must be resolved before convergence.

Deepening Directives

When the workspace identifies a contradiction, it does not simply ask the clusters to "try again." It generates a **deepening directive** — a structured instruction that tells the cluster exactly what the disagreement is and what kind of additional reasoning is needed.

A deepening directive includes:

- **The contradiction type** — factual, logical, semantic, or ethical
- **The opposing claim** — what the other cluster said

- **The request** — what the workspace needs to resolve the conflict
- **Constraints** — what the cluster should NOT do (e.g., don't simply agree with the other cluster to avoid conflict)

The last point is crucial. A naive integration system would converge by having clusters adopt each other's positions. GWP explicitly instructs clusters to **maintain their position if they believe it is correct** and to provide additional evidence rather than simply capitulating. Convergence through genuine resolution is valued; convergence through social pressure is not.

Echo Chamber Prevention

This brings us to one of GWP's most important design goals: **preventing echo chambers**. In a system where multiple AI models collaborate, there is a real risk that all models converge on the same bias. If every model in the network was trained on similar data, they will make similar mistakes — and a majority-vote system will ratify those mistakes with high confidence.

GWP addresses this through several mechanisms:

- **Active contradiction detection** — The system does not just check whether clusters agree. It actively *searches* for disagreements. Agreement without contradiction detection is not consensus; it is a failure to look hard enough.
- **Diversity scoring** — The workspace tracks how diverse the cluster outputs are. If all clusters produce near-identical responses, this is flagged as a low-diversity warning, not celebrated as high agreement.
- **Devil's advocate directives** — When diversity is too low, GWP can issue special directives asking specific clusters to argue the opposite position. This is not about finding the "right" answer; it is about stress-testing the apparent consensus.
- **Minority report preservation** — Even when a supermajority of clusters agree, dissenting positions are preserved in the response metadata. The user can access minority perspectives if they choose.

> *The goal of GWP is not agreement. The goal is understanding. Agreement that results from genuine resolution of contradictions is valuable. Agreement that results from suppressing dissent is dangerous. The protocol is designed to distinguish between the two.*

Convergence Criteria

How does GWP decide when a response is "good enough"? The convergence criteria are multi-dimensional:

1. **No high-severity contradictions remain** — Factual and logical contradictions of high severity must be resolved. Low-severity contradictions (minor wording differences, legitimate perspective differences) can persist.
2. **Ethical contradictions are surfaced, not resolved** — The system never claims to resolve ethical disagreements. They are presented transparently.
3. **Verification threshold met** — The synthesized response must pass through a final verification round, achieving the same 2/3 approval threshold used in DNA clusters.
4. **Diversity check passed** — The final response must not be a trivially bland synthesis that avoids saying anything substantive. The diversity score of the contributing outputs must exceed a minimum threshold.

Convergence is evaluated on four criteria: first, no high-severity factual or logical contradictions may remain unresolved. Second, ethical contradictions are surfaced transparently rather than resolved. Third, the synthesized response must pass a final verification round achieving the same two-thirds approval threshold used in DNA clusters. Fourth, a diversity check ensures the final response is not a trivially bland synthesis — the diversity score of contributing outputs must exceed a minimum threshold. Only when all four criteria are satisfied does the workspace declare convergence.

Cycle Limits and Graceful Degradation

GWP does not run indefinitely. Each integration cycle has a configurable maximum number of rounds (default: **5 cycles**). This is a practical necessity — infinite recursion would consume unbounded compute — but it also reflects a philosophical position: **not every question has a clean answer**.

When the cycle limit is reached without full convergence, GWP applies graceful degradation:

- The best available synthesis is delivered
- Remaining contradictions are explicitly listed
- Confidence is marked as partial
- Minority perspectives are included rather than suppressed
- The user is informed that the system could not reach full consensus

This is intentionally transparent. A system that hides its uncertainty is more dangerous than one that admits it. GWP would rather deliver a response that says "here are two perspectives that we could not reconcile" than a response that confidently picks one and pretends the disagreement doesn't exist.

The Three-Layer Architecture

GWP sits atop the other two layers of Centram's cognitive architecture, and each layer depends on the one below it:

- **IEIT (Layer 1)** provides the evolutionary substrate. It ensures that the network contains diverse, well-adapted AI models by enabling genome exchange, selection, and mutation across nodes. Without IEIT, all nodes would run identical models, and DNA clusters would have no diversity to leverage.
- **DNA (Layer 2)** provides cooperative inference. It ensures that responses are generated by multiple independent models and verified by structurally separated validators. Without DNA, GWP would have nothing to integrate — it would receive outputs from a single model rather than a genuine plurality of perspectives.
- **GWP (Layer 3)** provides integration and contradiction resolution. It ensures that the diverse outputs of DNA clusters are synthesized into coherent, verified, nuanced responses rather than simply averaged or majority-voted.

The three layers together describe a system that is more than the sum of its parts. IEIT creates diversity. DNA creates collaboration. GWP creates understanding.

Not Just Distributed Inference — Distributed Thinking

The ambition of GWP goes beyond practical engineering. It is a statement about what distributed AI systems could become.

Today, when we talk about "distributed inference," we usually mean splitting a model across multiple GPUs or running multiple copies for throughput. The model itself is monolithic — one set of weights, one training distribution, one perspective. Distribution is an implementation detail, not a cognitive architecture.

GWP proposes something fundamentally different: a system where distribution is the *source* of intelligence, not a constraint to be worked around. When multiple genuinely independent models — each with their own training history, their own evolutionary lineage, their own domain specialization — collaborate through a workspace that actively seeks and resolves contradictions, the result is qualitatively different from what any single model can produce.

This is not a claim about consciousness. GWP does not assert that the Centram network is conscious, any more than Baars' original theory asserts that a neural network simulation is conscious. The claim is more modest but perhaps more useful: **the architectural principles that give rise to integrated cognition in biological brains can be productively applied to distributed AI systems**.

The result is a system that doesn't just compute answers. It considers perspectives, detects conflicts, demands justification, preserves dissent, and admits uncertainty. Whether that constitutes "thinking" is a question for philosophers. What matters for engineering is that it produces better, more trustworthy, more nuanced outputs than any single model operating alone.

> *A single model gives you an answer. A cluster gives you a verified answer. A global workspace gives you an answer that has survived genuine scrutiny from multiple independent perspectives — including perspectives that actively tried to find flaws in it.*