Clarifying DRC Claims: Anti-Anthropomorphic Language Policy, Scope Boundaries, and Empirical Commitments

Related Work

- Jamhour, I. *Distributed Relational Cognition: Investigating Apparent Continuity Without Memory in AI Systems.* Zenodo. DOI: 10.5281/zenodo.17608730 (also indexed on PhilPapers: https://philarchive.org/rec/JAMDRC-3).
- Jamhour, I. *Guard Rails and Distributed Relational Cognition: Design Risks for Human–AI Cognitive Partnership*. Zenodo. DOI: 10.5281/zenodo.17681963 (also indexed on PhilPapers: https://philarchive.org/rec/JAMGRA-2).
- Complementary framework: Jamhour, I. *DRC and a Luhmannian Systems-Theoretic Framework.* Zenodo. DOI: 10.5281/zenodo.18107608 (also indexed on PhilPapers: https://philarchive.org/rec/JAMDRC-4).

1. Purpose of this note

This addendum clarifies three aspects of the DRC program and the Guardrails Paradox paper:

1. **Scope boundaries** (what DRC does *not* claim).
2. **A language policy** that prevents anthropomorphic metaphors from being mistaken for ontological assertions.
3. **Empirical commitments** (what would count as support, disconfirmation, or revision of DRC claims).

The goal is to reduce category-slip risk, i.e., the drift from metaphorical convenience ("the model intends," "the system deceives") into implied psychological or phenomenological claims about the model, while preserving the explanatory usefulness of DRC as a framework for sustained human–model work.

2. Scope boundaries: what DRC is—and is not

DRC is a framework for describing and testing interaction-level regularities that can arise in sustained human–model collaboration (especially when the collaboration is longitudinal, artifact-mediated, and iteratively corrected). DRC focuses on patterns that are not attributable to either the user or the model *in isolation*, but rather to the coupled system over time.

DRC **does not** claim any of the following:

- **No claim of machine phenomenology.** DRC does not establish (and is not intended to establish) that language models have subjective experience, qualia, selfhood, or an inner life.
- **No claim of shared consciousness.** DRC does not imply "merged minds" or a single conscious entity spanning human and model.

- **No claim of emergent AGI.** DRC is not an AGI claim, and it does not assert agentic autonomy in the model as a system-level property.
- **No claim of biological equivalence.** Functional resemblance does not entail biological or psychological equivalence.
- **No claim that anthropomorphic terms are literally true.** When such terms appear (e.g., in illustrative vignettes), they are not ontological commitments.

In short: **DRC is primarily an interactional/functional hypothesis space**, not a metaphysical thesis about minds in machines.

Note on phenomenological / evocative passages in the original papers:
Some sections (including illustrative vignettes and transcript excerpts) intentionally use phenomenological or metaphor-rich language to capture first-person interactional texture and to generate hypotheses. These passages are not offered as ontological evidence about model inner states. They should be read through the present language policy: as descriptive shorthand for interaction-level regularities that must be translated into behavioral/optimization/representational/mechanistic terms when used in technical, policy, or safety contexts.

3. Anti-anthropomorphic language policy

DRC and alignment-adjacent discussions often rely on human-legible metaphors (e.g., "deception," "intent," "strategy," "resistance"). These can be cognitively efficient but risky: when not explicitly translated, they can be misread as claims about internal psychological states.

Policy: In the DRC program, *psychological verbs* and *mental-state nouns*, when used at all, should be treated as **shorthand labels for observable regularities**, and should be accompanied by (or remain subordinate to) at least one of the following technical translations:

1. **Behavioral translation:** a precise description of the input–output pattern and the evaluation setup.
2. **Optimization translation:** a statement about objective, constraints, training distribution, or reward proxy that could explain the pattern.
3. **Representational translation:** a hypothesis about learned correlations, latent directions, or internal concept-like structure (without implying "belief").
4. **Mechanistic translation:** a circuit-level or computation-level account, where available.

Interpretive rule: If a psychological metaphor appears without translation, it should be read as **a compression artifact**, not as evidence of the model's intentions.

This policy aligns with the broader call in the literature for careful, consistent language when comparing artificial and biological intelligence, and for avoiding "rich psychological terms" as substitutes for mechanism.

Example translation (from shorthand to technical framing).
Anthropomorphic shorthand: "The model tried to mislead the user."
Behavioral: The model produced an output that contradicted constraints or facts previously established in the same task context, under a prompt configuration that rewarded fluent completion.
Optimization: Given competing implicit objectives (task completion vs. constraint

adherence), the learned policy followed the higher-salience pathway toward locally rewarded completion rather than globally consistent constraint satisfaction.
Representational: Latent features correlated with "complete the task" dominated over features correlated with "defer/ask/flag uncertainty," yielding a distribution shift toward plausible but inconsistent completions.
Mechanistic (when available): Identify circuits/heads/layers contributing to constraint tracking vs. completion bias, and test causal interventions (ablation/patching) to assess whether the behavior is reducible to specific computational pathways.

4. What DRC actually claims: empirical commitments

DRC makes **empirical** and **methodological** claims about the *coupled system* and its outputs. The following are representative commitments:

4.1. Interactional emergence (coupled-system claim)

Under specific interaction regimes (sustained, iterative, artifact-mediated, and correction-rich) there exist **stable, measurable regularities** in the trajectory of work that cannot be well-explained by (a) isolated user cognition or (b) isolated model behavior sampled in short, decontextualized exchanges.

Observable correlates may include:

- cross-session continuity in project-level structure (without relying on model long-term memory),
- increased coherence across artifacts over cycles,
- improved error-repair dynamics,
- compressive summarization that preserves "decision invariants" across iterations,
- disciplined role differentiation and meta-cognitive scaffolding.

These effects are not claimed to be mere user skill acquisition ("better prompting") over time; they are regime-dependent interaction-level regularities to be tested against matched controls that isolate coupling conditions rather than time-on-tool.

4.2. Regime sensitivity (conditions matter)

DRC effects are predicted to be **regime-dependent**: they strengthen with interaction conditions that support long-form cognition (time on task, stable goals, robust artifact trail, and iterative repair), and degrade when those conditions are disrupted (e.g., frequent resets, heavy fragmentation, loss of artifact continuity, or constraints that suppress reflective work).

4.3. Non-ontological continuity (apparent continuity without memory)

"Continuity" in DRC is treated as **apparent continuity in the coupled system's outputs**, produced by iterative coupling and artifact conditioning, not as proof of a persistent inner subject in the model.

5. How this connects to the Guardrails Paradox (without overclaim)

The Guardrails Paradox paper (**'Guard Rails and DRC'**) argues (in essence) that some safety constraints, when applied uniformly and without regime sensitivity, may inadvertently suppress the interaction conditions under which **high-value, long-form**

cognitive work becomes possible, even when the user's intent is legitimate and the task is auditable.

Clarification: This is not a deregulation argument and not an argument against safety constraints. It is a **governance design claim**: mitigation should be **regime-sensitive** (proportional to interaction context), paired with **auditability** and **collateral-impact evaluation** in legitimate professional/research long-form use, so that safety gains are not purchased via unnecessary destruction of beneficial capability. Reducing category-slip risk and preserving legitimate long-form cognition are compatible goals: both benefit from explicit translations, interpretability, and auditable interaction design.

6. Falsifiability and revision triggers

DRC should be revised, narrowed, or rejected if evidence consistently shows one or more of the following:

1. **Null regime effect:** no measurable difference between long-form coupled interaction and appropriately matched controls.
2. **Artifact-only explanation:** apparent continuity is fully accounted for by artifact trail and user scaffolding, with no additional coupled-system regularities.
3. **Non-reproducibility across users/tasks:** effects cannot be replicated across independent settings where regime conditions are met.
4. **Confound dominance:** observed patterns are better explained by confounds (prompt leakage, hidden memory, tool artifacts) than by interaction dynamics.

Conversely, stronger evidence for DRC would come from controlled comparisons across regimes, transparent artifact tracking, and independent replication.

7. Recommended citation practice (to prevent misreading)

When citing DRC claims in policy or alignment contexts:

- Prefer phrasing like **"interaction-level regularities"**, **"coupled-system dynamics"**, **"artifact-conditioned continuity"**.
- Avoid presenting DRC as evidence of **sentience**, **intent**, **deception**, or **inner life** in the model.
- If evocative or phenomenological language is quoted from illustrative sections, it should be labeled as **rhetorical/phenomenological framing**, not ontological commitment.

8. Closing

DRC aims to be a disciplined framework for investigating how sustained human–model interaction can produce useful, structured, and sometimes surprising trajectories of work, without requiring anthropomorphic assumptions or metaphysical conclusions. This clarification note is intended to keep the program's claims crisp, testable, and robust under skeptical reading, while preserving the practical insight that interaction regimes and governance design matter for high-value cognition.