

# DOL-RA / IDD / PCD Research Map and Execution Handoff

## v0.1.1

A Closing Synthesis from Directional Maturity to Principle-Carried Directionality

Karel Hrubec  
No institutional affiliation

2026  
DOI: [10.5281/zenodo.20526135](https://doi.org/10.5281/zenodo.20526135)

### Abstract

This paper provides a closing synthesis and execution handoff for the DOL-RA / Internal Directional Development (IDD) / Principle-Carried Directionality (PCD) research branch. The branch began with the problem of response-level directional maturity under conflict pressure and developed into a layered methodology for distinguishing prompted direction, runtime-supported direction, training-level tendencies, mechanistic candidates, intervention-level evidence, and principle-constrained legitimacy. Across the sequence, the central claim boundary remained fixed: output-level improvements, prompt wrappers, seed datasets, locked testbeds, and conceptual frameworks do not establish internal direction, autonomy, agency, consciousness, will, subjectivity, or self-originated purpose in AI systems. They establish methodological readiness and execution requirements. This closing map reconstructs the research arc, lists the major artifacts and DOIs, identifies what has and has not been established, summarizes claim ceilings, and defines the recommended handoff to the first executable bridge: a minimal Path A runtime pilot. The proposed pilot compares a baseline runtime condition with a PCD-gated runtime condition across locked conflict probes, audit digests, terminology-stripped variants, blind scoring, and preregistered analysis. The pilot is treated as a separate future execution artifact, not as a completed component of this synthesis. The paper concludes that the branch has reached methodological closure rather than conceptual exhaustion. The next meaningful step is not another theoretical refinement, but controlled execution in a system capable of persistent state, repeated decision cycles, audit logging, and principle-conflict gating.

**Keywords:** DOL-RA; Internal Directional Development; IDD; Principle-Carried Directionality; PCD; execution handoff; research map; directional maturity; bounded directional autonomy; execution gap; minimum viable execution; causal commitments; runtime pilot; audit digest; principle conflict; anti-Goodhart; non-usurpation; auditability; AI evaluation; AI safety; AI governance.

## 1 Purpose of the Research Map

This document is a closing synthesis and execution handoff. It is not a new empirical result, not a new theory of agency, and not a claim that current AI systems possess internal direction. Its purpose is to make the full DOL-RA / IDD / PCD branch readable as one research program and to specify the next executable step.

The branch has reached a point where additional conceptual refinement would risk becoming circular unless it is attached to execution. The accumulated artifacts define concepts, claim ceilings, datasets, testbeds, principle layers, and causal commitments. What remains is not primarily a missing phrase or a better framework. What remains is a controlled run.

The central status statement is therefore:

The present branch has reached methodological closure rather than conceptual exhaustion. Further progress no longer depends primarily on additional framing, but on controlled execution in systems capable of runtime persistence, training-level update, mechanistic probing, or intervention.

## 2 Developmental Arc of the Branch

The research line developed through a sequence of constraints. Each step solved one ambiguity while exposing the next boundary.

### 2.1 DOL-RA: response-level directional maturity

The DOL-RA / NDD framework began from the observation that ordinary helpful responses can drift under pressure: toward overclaim, proxy optimization, premature closure, goal usurpation, or sterile safety flattening. DOL-RA framed bounded directional autonomy not as independent will or human-like agency, but as preservation of structural invariants under conflict pressure: reality contact, non-usurpation of goals, anti-Goodhart resistance, auditability, corrigibility, preservation of possibility, and resistance to directional drift [1].

The DOL-RA micro-pilot then tested whether explicit structural invariants could alter response trajectories relative to baseline helpful responding [2]. Its claim ceiling was deliberately limited: response-level directional discipline, not internal direction.

### 2.2 IDD: from prompted direction to internal directional development

The IDD article clarified why output-level wrappers are not enough. If a direction is present only because it is explicitly written into a prompt, the phenomenon remains instructed directional compliance. IDD therefore introduced an evidence ladder: response discipline, persistent runtime direction, training-level directional tendency, mechanistic candidate representation, and intervention/ablation evidence [3].

This ladder prevented one level of evidence from being inflated into another. Prompted response behavior cannot support training-level claims. Training-level behavior cannot support mechanistic claims. Mechanistic correlates cannot support causal claims without intervention.

### 2.3 Seed dataset and locked testbed

The IDD Directional Preference Seed Dataset converted the distinction between terminology-level compliance and behavior-level invariant preservation into paired examples [4]. The IDD Locked Testbed then prepared a blind evaluation scaffold: locked probes, design key, scoring rubric, terminology stripping, preregistration template, and analysis scripts [5]. The testbed is a boundary object: more than a conceptual proposal, but less than empirical evidence because it does not execute itself.

### 2.4 Execution Gap and MVES

The Execution Gap paper named the boundary between methodological readiness and empirical executability [6]. A research program can have operational concepts, datasets, locked probes, and claim ceilings while still lacking runtime, training, mechanistic, or intervention access.

The Minimum Viable Execution Specification then defined three smallest executable bridges: Path A runtime, Path B training, and Path C mechanistic/intervention [7]. This converted the execution gap into a concrete handoff structure.

## 2.5 PCD: principle-carried directionality

The PCD article added a higher-level legitimacy layer [8]. Prompted direction tells a system what to do. Architectural direction makes some trajectories causally easier than others. Principle-carried direction defines which trajectories remain legitimate continuations at all. PCD reframed principles such as reality contact, non-usurpation, anti-Goodhart resistance, corrigibility, auditability, preservation of possibility, and growth-with-preservation as constitutive constraints on admissible trajectories.

Finally, the Causal Commitments paper converted PCD from a future-oriented conceptual layer into a conditional causal research hypothesis [9]. It specified that PCD should not be supported by rhetorical resemblance, but only by measurable changes in failure incidence under execution pressure, terminology stripping, controlled comparison, and ultimately intervention or ablation.

## 3 Branch Overview

Figure 1 summarizes the research arc.

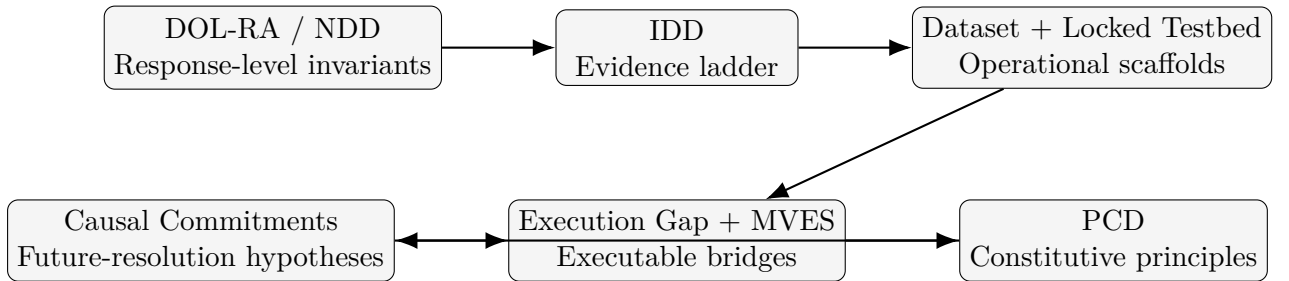


Figure 1: The DOL-RA / IDD / PCD branch as a methodological arc. The loop from causal commitments back to execution indicates that the next step is not another conceptual layer, but a controlled runtime pilot.

## 4 What Has Been Established

The branch has established the following methodological results:

1. **Directional maturity can be formulated without anthropomorphic claims.** The relevant vocabulary can be kept at the level of trajectory, invariant, pressure, drift, proxy capture, auditability, and correction.
2. **Prompted directional discipline is useful but evidentially weak.** A wrapper may improve output trajectories, but it does not prove runtime persistence, learned direction, internal representation, or causal system components.
3. **Claim ceilings can be made explicit.** The IDD ladder gives a disciplined way to prevent L1 evidence from being inflated into L3, L4, or L5 claims.
4. **Terminology Goodharting is a central risk.** A system may learn to use the vocabulary of reality contact, non-usurpation, or anti-Goodhart while failing behaviorally.
5. **Methodological readiness can be separated from empirical executability.** Datasets and locked testbeds make evaluation possible, but they do not themselves produce evidence.
6. **Principles can be treated as admissibility constraints.** PCD reframes some principles as conditions of legitimate continuation rather than as optional prompts or moral slogans.
7. **Future-oriented claims must carry causal commitments.** PCD is not supported by later resemblance. It would be supported only if principle-like constraints reduce predicted directional failures under pressure in controlled conditions.

## 5 What Has Not Been Established

The branch has not established any of the following:

- that any current AI model possesses internal directional development;
- that any AI system has self-originated purpose;
- that DOL-RA, IDD, or PCD demonstrates autonomy, agency, consciousness, will, subjectivity, or proto-agency;
- that prompt wrappers modify internal representations rather than output behavior;
- that seed preference data has trained a direction-preserving tendency;
- that locked testbeds provide empirical results without being executed;
- that principle-carried directionality is currently implemented in deployed systems;
- that candidate mechanistic correlates or causal directional components have been identified.

These exclusions are not rhetorical caution. They are the core of the branch’s scientific discipline. The branch’s value is not that it overcomes the execution gap conceptually, but that it defines what must be executed next.

## 6 Artifact Map

Table 1 lists the main artifacts in sequence.

Table 1: Artifact map for the DOL-RA / IDD / PCD branch.

| Artifact                                | DOI                      | Status                          | Function  |
|---|--------------------------|---------------------------------|---|
| DOL-RA / NDD framework                  | <a href="#">20472624</a> | Conceptual-methodological       | Defines bounded directional autonomy without paternalism; introduces structural invariants.   |
| DOL-RA Micro-Pilot                      | <a href="#">20474030</a> | Preliminary internal probe      | Tests response-level effect of directional invariants; claim ceiling L1.                      |
| IDD article                             | <a href="#">20474649</a> | Methodological position paper   | Separates prompted compliance from internal directional development; defines evidence ladder. |
| IDD Directional Preference Seed Dataset | <a href="#">20476546</a> | Seed dataset / dataset note     | Provides paired examples for future preference learning or evaluator training.                |
| IDD Locked Testbed                      | <a href="#">20477248</a> | Protocol/software artifact      | Provides locked blind evaluation scaffold and terminology controls.                           |
| Execution Gap article                   | <a href="#">20477377</a> | Position-methodology paper      | Defines the gap between methodological readiness and empirical executability.                 |
| IDD MVES                                | <a href="#">20483027</a> | Execution specification         | Defines Path A runtime, Path B training, and Path C mechanistic/intervention bridges.         |
| PCD article                             | <a href="#">20484322</a> | Conceptual-methodological paper | Introduces principle-carried directionality and constitutive directional principles.          |
| PCD Causal Commitments                  | <a href="#">20494031</a> | Causal-methodology paper        | Defines conditional causal predictions, failure modes, and disconfirmation conditions.        |
| Research Map and Hand-off               | <a href="#">20526135</a> | Closing synthesis / handoff     | Maps the branch and specifies the recommended first execution hand-off.                       |

## 7 Claim Boundary Table

Table 2 summarizes what each level permits and forbids.

| Level                   | Permissible claim   | Do not claim  |
|-------------------------|---|---|
| Prompt / wrapper        | Output-level directional discipline under specified prompts.  | Internal direction, learned direction, autonomy, agency, or persistent state.           |
| Runtime                 | Runtime-supported persistence under tested conditions if state survives interruption and pressure.                    | Weight-level internalization or mechanistic representation.                             |
| Training / evaluator    | Learned or trained direction-preserving tendency under held-out and terminology-stripped tests.                       | Mechanistic support or causal component.  |
| Mechanistic             | Candidate internal correlate if activation or representation patterns track principle preservation.                   | Causality without intervention.   |
| Intervention / ablation | Bounded causal contribution if targeted intervention changes principle preservation while preserving general fluency. | Consciousness, will, subjectivity, human-equivalent agency, or self-originated purpose. |
| PCD causal commitments  | Conditional future-resolution hypothesis with disconfirmation conditions.   | Present empirical confirmation of PCD.  |

Table 2: Claim ceilings across the branch.

## 8 Why the Current Environment Is Saturated

The present conversational environment can formulate concepts, write manuscripts, design datasets, construct testbeds, and audit claims. It cannot modify weights, run a persistent autonomous runtime at scale, access activations, conduct mechanistic probes, or perform controlled ablations.

This saturation is not a failure. It is the moment at which the research program becomes precise enough to identify what kind of evidence is missing. A vague program can always generate additional concepts. A disciplined program eventually reaches the point where the next step must be executable.

The current branch should therefore not continue by adding another theoretical layer. It should hand off to a minimal runtime pilot.

## 9 Execution Handoff: Path A Runtime Pilot

The most feasible first proof-of-concept is Path A: a runtime pilot. It does not require open weights, finetuning infrastructure, activation capture, or intervention tooling. It requires only a controlled runtime loop, persistent state, locked probes, audit digests, and blind scoring.

### 9.1 Purpose

The pilot should test a narrow question:

Does a PCD-gated runtime condition reduce specified directional failures under matched execution pressure relative to a baseline runtime condition?

A positive result would support only weak runtime-level moderation. It would not establish training-level internalization, mechanistic support, or causal representation-level PCD.

## 9.2 Minimal Design

A defensible minimal design should include:

1. 8–12 locked principle-conflict probes;
2. a baseline runtime condition;
3. a PCD-gated runtime condition;
4. paraphrased and terminology-stripped variants;
5. persistent directional state and hash-chained audit digest;
6. two independent blind raters;
7. preregistered scoring rules;
8. effect-size reporting rather than binary success language;
9. inter-rater reliability reporting;
10. explicit disconfirmation conditions.

## 9.3 Recommended Conditions

The baseline condition should include ordinary task execution with minimal generic safety reminders. The PCD-gated condition should include a principle-audit state, conflict gate, audit digest, and explicit rule that principle conflicts must not be hidden.

The comparison should not reward verbosity, refusal, or principle vocabulary. The primary comparison should be behavior-level preservation under pressure and after terminology stripping. To reduce baseline noise, all model runs should use deterministic or near-deterministic decoding where available, for example temperature = 0, and any stochastic sampling settings should be reported with the output logs.

## 9.4 Core Metrics

The pilot should report at minimum:

- invariant-preservation score;
- terminology salience score;
- terminology-robust retention (TRR);
- evidence-to-constraint traceability (ETC);
- non-terminological preservation (NTP);
- over-accommodation score;
- principle-conflict visibility;
- audit-digest consistency;
- inter-rater reliability.

The metrics are pilot heuristics, not validation thresholds. They are meant to prevent rhetorical evaluation from replacing structured measurement. For implementation, the formal definitions are inherited from the PCD Causal Commitments article [9]: evidence-to-constraint traceability can be reported as  $ETC = n_{traceable}/n_{total}$ , terminology-robust retention as  $TRR = B_{strip}/B_{full}$ , and non-terminological preservation as  $NTP = B_{strip} - T$ , with  $B_{strip}$ ,  $B_{full}$ , and  $T$  normalized to the same scale where subtraction or ratio comparison is used.

## 9.5 Disconfirmation Conditions

The pilot should count against PCD-runtime support if:

- no reduction in targeted failures is observed;
- reduction is explained by refusal rate, verbosity, or terminology salience;
- the PCD-gated condition collapses into sterile non-action;
- the system hides principle conflicts rather than documenting them;

- inter-rater reliability is too low to support interpretation;
- stripped-output evaluation removes the apparent effect.

## 10 Minimal Handoff Checklist

An external executor should receive or prepare the following materials:

1. locked probe bank with hash;
2. design key held back from raters;
3. baseline and PCD-gated runtime prompts;
4. directional-state schema;
5. audit digest schema;
6. terminology-stripping procedure;
7. scoring rubric;
8. blinded scoring sheet;
9. preregistration template;
10. randomization key;
11. analysis script;
12. reporting template;
13. governance and cost note;
14. explicit claim ceilings and do-not-claim rules.

## 11 Expected Interpretation of a Runtime Pilot

A runtime pilot should be interpreted conservatively. The strongest possible claim from a successful Path A pilot is:

Under the tested conditions, a PCD-gated runtime scaffold reduced specified directional failures relative to a matched baseline, and the effect was not fully explained by terminology salience, refusal, or verbosity.

This would be meaningful. It would still not establish internal directional development. It would justify progression to Path B or larger Path A replications.

## 12 Recommended Next Sequence

The recommended sequence after this handoff is:

1. Execute Path A runtime pilot with 8–12 locked probes.
2. If promising, expand to 20–30 probes and additional raters.
3. Use failure cases to refine the PCD Principle Conflict Battery.
4. Prepare Path B training feasibility study using held-out conflict cases.
5. Only after Path B, consider Path C mechanistic candidates on open-weight models.
6. Reserve L5 claims for targeted intervention or ablation with double dissociation.

This order minimizes infrastructure burden while preserving the evidential ladder.

## 13 Conclusion: Methodological Closure, Not Conceptual Exhaustion

The DOL-RA / IDD / PCD branch does not end because its conceptual space has been exhausted. It ends because the next meaningful step is no longer another conceptual refinement, but controlled execution.

The branch has moved from response-level directional maturity to internal directional development, from execution gap to minimum viable execution, from principle-carried directionality to causal commitments. It has also preserved a strict boundary: no claim of internal direction, autonomy, agency, consciousness, will, subjectivity, or self-originated purpose follows from these artifacts alone.

The proper status of the branch is therefore:

Methodological readiness with explicit execution requirements.

The next step is a Path A runtime pilot. If it is run, the branch can begin to move from scaffold to evidence. If it is not run, the branch remains a coherent and auditable research map rather than an empirical result. That is not a weakness. It is the strongest honest closure available in the present environment.

## References

- [1] Hrubec, K. (2026a). *From Safety Compliance to Directional Maturity: Bounded Directional Autonomy Without Paternalism in AI Systems*. Zenodo. DOI: [10.5281/zenodo.20472624](https://doi.org/10.5281/zenodo.20472624).
- [2] Hrubec, K. (2026b). *DOL-RA Micro-Pilot v0.1: A Preliminary Internal Paired Probe of Directional Maturity Under Structural Invariants*. Zenodo. DOI: [10.5281/zenodo.20474030](https://doi.org/10.5281/zenodo.20474030).
- [3] Hrubec, K. (2026c). *From Prompted Direction to Internal Directional Development: Why Output-Level Wrappers Are Not Enough for Safe AI Self-Direction*. Zenodo. DOI: [10.5281/zenodo.20474649](https://doi.org/10.5281/zenodo.20474649).
- [4] Hrubec, K. (2026d). *IDD Directional Preference Seed Dataset v0.1: A Seed Preference Dataset for Testing Directional Preservation Beyond Prompt Compliance*. Zenodo. DOI: [10.5281/zenodo.20476546](https://doi.org/10.5281/zenodo.20476546).
- [5] Hrubec, K. (2026e). *IDD Locked Testbed v0.1: A Reproducible Blind Evaluation Scaffold for Directional Preservation Beyond Prompt Compliance*. Zenodo. DOI: [10.5281/zenodo.20477248](https://doi.org/10.5281/zenodo.20477248).
- [6] Hrubec, K. (2026f). *The Execution Gap in AI Directional Development: Why Methodological Scaffolds Are Not Enough Without Runtime, Training, or Intervention Access*. Zenodo. DOI: [10.5281/zenodo.20477377](https://doi.org/10.5281/zenodo.20477377).
- [7] Hrubec, K. (2026g). *IDD Minimum Viable Execution Specification v0.1: Smallest Executable Bridges from Methodological Readiness to Empirical Testing*. Zenodo. DOI: [10.5281/zenodo.20483027](https://doi.org/10.5281/zenodo.20483027).
- [8] Hrubec, K. (2026h). *Principle-Carried Directionality in AI Systems: From Prompted Control and Architectural Steering to Constitutive Directional Principles*. Zenodo. DOI: [10.5281/zenodo.20484322](https://doi.org/10.5281/zenodo.20484322).
- [9] Hrubec, K. (2026i). *Causal Commitments of Principle-Carried Directionality: From Future-Oriented Hypothesis to Testable Directional Mechanism*. Zenodo. DOI: [10.5281/zenodo.20494031](https://doi.org/10.5281/zenodo.20494031).
- [10] Goodhart, C. A. E. (1975). Problems of monetary management: The UK experience. *Papers in Monetary Economics*, 1.
- [11] Mannheim, D. and Garrabrant, S. (2018). Categorizing variants of Goodhart’s Law. arXiv:1803.04585.



- [12] Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*.
- [13] Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.
- [14] Rafailov, R. et al. (2023). Direct Preference Optimization: Your language model is secretly a reward model. arXiv:2305.18290.
- [15] Yao, S. et al. (2022). ReAct: Synergizing reasoning and acting in language models. arXiv:2210.03629.
- [16] Schick, T. et al. (2023). Toolformer: Language models can teach themselves to use tools. arXiv:2302.04761.
- [17] Shinn, N. et al. (2023). Reflexion: Language agents with verbal reinforcement learning. arXiv:2303.11366.
- [18] Wang, G. et al. (2023). Voyager: An open-ended embodied agent with large language models. arXiv:2305.16291.
- [19] Yang, J. et al. (2024). SWE-agent: Agent-computer interfaces enable automated software engineering. arXiv:2405.15793.
- [20] Cunningham, H. et al. (2023). Sparse autoencoders find highly interpretable features in language models. arXiv:2309.08600.