# Evidence and Open Questions

## How Normative Detachment Theory Could Be Studied

**Bree Beal Version 1.10 - September 2026 Status:** Unratified RC4 author draft; not an approved theory artifact.

Normative Detachment Theory (NDT) proposes an account of how someone forms normative judgments and of what makes some of those judgment events moral. It is not one empirical hypothesis awaiting one decisive experiment. It is a layered theory whose claims call for different kinds of evaluation.

Some questions are conceptual: Does the theory identify a coherent kind without defining moral cognition by moral words, approved answers, or familiar behaviors? Some are comparative: Does its proposed structure preserve important distinctions better than rival accounts? Some are empirical: Can that structure be identified in actual cognitive episodes? Others concern measurement: Can its components be measured with enough reliability and precision to support quantitative predictions?

Keeping these questions separate makes the theory more answerable to evidence, not less. It also prevents a successful local study from being mistaken for confirmation of the whole theory, or an immature measurement model from being mistaken for the absence of a genuine theoretical distinction.

This document explains what is fixed in the current theory, what remains open, and what several possible research programs could establish.

---

## 1. Four standards that should not be collapsed

NDT distinguishes four standards of assessment. A constitutive role is a claim type—what the theory proposes as part of the kind—not a fifth evidential level.

**Standard**: **Comparative explication adequacy**

**Plain-language question**: Does NDT characterize the target kind better than alternatives?

**Appropriate evidence**: Noncircularity; paradigm and contrast coverage; explanatory integration and discrimination; independently specified consequences

---

**Standard**: **Formal countermodel structure**

**Plain-language question**: Do the definitions generate explicit structures in which registered conclusions hold or fail?

**Appropriate evidence**: Finite countermodels; dependency checks; theorem and corollary status; executable conformance where encoded

---

**Standard**: **Operational identification**

**Plain-language question**: Can real episodes instantiating the organization be distinguished from plausible rivals?

**Appropriate evidence**: Independent indicators; pre-specified interventions, transfer, and process evidence; held-out comparison

---

**Standard**: **Calibrated measurement**

**Plain-language question**: Can components receive validated scores, weights, thresholds, and uncertainty in a declared range?

**Appropriate evidence**: Reliability, validity, calibration, invariance, uncertainty, and out-of-sample prediction

A theory can be comparatively and formally precise while remaining operationally or quantitatively incomplete. That is NDT's present position. Constitutive status does not immunize an explication from assessment: sustained failure on independently specified contrast or consequence tests can require restriction, revision, or abandonment. Where adequate measurement is not yet available, that is a registered limitation rather than evidence in the theory's favor.

The distinction matters especially when ordinary uses of *moral* vary. An alleged counterexample such as an "agentless moral judgment" cannot simply be counted against NDT as if everyone had already agreed on the classification that is at issue. A lexical verdict cannot settle an agentless case; §§6B and 8 state when independently measured evidence becomes adverse.

---

# 2. What the current theory proposes

## Normative judgment

NDT uses **detachment** in the deontic sense: someone forms guidance or appraisal from a represented action-guiding situation. Ordinary normative thought includes permissions, but the current six-output formal grammar does not yet contain primitive permission as its own output type. Explicit permission, authorization, exception, and protected choice remain extension questions. The theory explains judgment formation, not whether the judgment is later enacted.

Someone forming a normative judgment represents such things as possible actions, likely outcomes, standards, requirements, priorities, feasibility, and challenges. NDT takes those representations and inferential resources as supplied inputs; it does not yet model their end-to-end online generation. A normative conclusion is not derived from a description stipulated to be empty of every relevant evaluation, goal, ordering, or commitment. The broader evaluative prerequisite says that some operative evaluation must already be present, attributed, encoded, or derived. Constitutive evaluativity is the narrower claim that a representation can be evaluative partly because of its own content or functional role. A decision model must still use the operative evaluation to construct guidance or appraisal.

## Moral judgment

Moral cognition is a structurally defined subset of normative cognition. In the current account, a moral judgment includes:

1. a represented subject and object under an active interpretation;
2. a relationship core with a nonempty inherent-valence profile, Relational Positioning, and perceived agency;
3. the appropriate first-person or third-person subject condition; and
4. actual contribution of the bound relationship, subject, and object organization to the evaluation used in this occurrence of judgment.

**Inherent valence** has independent positive and negative components. Positive-only, negative-only, mixed, and neither profiles remain distinguishable; opposite signs do not cancel. **Relational Positioning** locates the exact object relative to the exact subject through identity-mediated positioning, intimacy, or direct encounter. Identity-mediated forms are shared, contrastive, and complementary-role. Mere category membership, role possession, physical proximity, or abstract recognition is insufficient. **Perceived agency** represents this subject as able to affect, respond to, answer to, or treat this object within a declared horizon.

The finite companion derives separate positive-valence-by-agency and negative-valence-by-agency branches before a profile-preserving relationship-strength gate, which is a derived application gate rather than a fourth relationship builder.

There is no universal numerical cutoff. The finite companion uses a replaceable component gate for explicit tests, not a population-invariant measurement claim. Functional access belongs to contribution machinery, not the relationship core. A route must be available to the process that uses it, and the bound profile must actually matter on the judgment-producing chain.

A **strict bender** is also separate: it preserves the exact relationship-core snapshot while acting at one fixed downstream locus through the required live contributing path. If a factor changes inherent valence, positioning, or perceived agency, it is a builder input, realizer, or modulator instead. The same representation may support several typed edges, but each edge must be identified and credited independently.

---

# 3. Why the relationship roles are not ordinary empirical predictors

The three core roles are not introduced as variables that happened to correlate with responses already called moral. They state different jobs in NDT's current explication.

| Role removed | Distinction the account can no longer preserve |
|---|---|
| **Inherent valence** | Inherent positive or negative significance versus merely instrumental relevance; deleting one component also tests positive-only, negative-only, and mixed profiles without cancellation |

| Role removed | Distinction the account can no longer preserve |
|---|---|
| **Relational Positioning** | Abstract evaluation of an object versus locating this exact object relative to this exact subject |
| **Perceived agency** | Passive appraisal versus representing this subject as able to affect, answer to, or respond to this object |

Route and edge lesions answer narrower questions. Deleting shared, contrastive, or complementary-role identity tests one identity-mediated form. Deleting intimacy or direct encounter tests one alternative route. Blocking access tests whether an intact route can enter the operative process. Deleting a strict bender while holding the core fixed tests the predicted downstream change. None of those lesions is a substitute for deleting a whole core role.

These contrasts do not establish unique terminology, validated indicators, or a universal gate. They define what a measure or rival must discriminate. A study that defines a role by the moral response it predicts would merely redescribe the outcome.

# 4. What remains genuinely open

The current architecture is a typed skeleton, not a finished psychometric instrument.

## Indicators

What observable evidence identifies positive and negative inherent value, each positioning route and identity form, perceived agency, functional access, strict benders, and profile-sensitive actual contribution without inferring them from the target judgment?

## Route and form discrimination

Can shared, contrastive, and complementary-role identity-mediated positioning be distinguished from one another and from intimacy and direct encounter? Can an apparently "distant" case be resolved as a missing route, a complementary-role bridge, or a downstream bender without reviving standing-mediated positioning as a catch-all?

## Valence profiles and quantitative form

Can positive-only, negative-only, mixed, and neither profiles be identified independently? Which component gates, interactions, or probabilistic rules work in a declared population? The current finite disjunction is replaceable; no universal numerical threshold is asserted.

## Bender boundary

Can a candidate factor be varied while the relationship core remains fixed, and does it change a pre-specified downstream locus? Power, vulnerability, sentience, authority, urgency, and institutional status are candidate sources. They are not automatically benders, and the 2021 discussion does not by itself establish their current edge type.

## Actual contribution under redundant support

Several routes may support the same output. Simple but-for dependence can be too demanding, while activation or reachability is too permissive. Identification must distinguish an available backup from a contributor and must not reuse evidence from one route or edge to certify another.

## Population, implementation, and kind boundary

Indicators and implementations may vary by population, task, and frame. Their operating range must be declared. Disputed uses of *moral* require comparative explication alongside measurement; lexical agreement alone cannot settle the kind.

# 5. Principles for a credible study

Whatever part of NDT is studied, five safeguards are essential.

## 1. Identify the proposed cause before the outcome

A relationship role, frame, challenge, positioning route or identity form, bender, or control state cannot be defined by the judgment or behavior it is supposed to predict. Evidence may corroborate an antecedent measure, but it cannot supply its definition after the fact.

### 2. Separate the judgment from its expression and enactment

What a participant says, chooses, or does can be evidence about a judgment. It is not identical to that judgment. Action also depends on motivation, ability, habit, coercion, and circumstance.

### 3. Compare plausible rivals

A study should ask whether the data favor relationship provenance, free-standing policy use, outcome-only choice, associative learning, demand compliance, character appraisal, or another specified alternative. Finding that one NDT-shaped model fits is less informative if obvious rivals were never tested.

### 4. Fix the comparison before inspecting the target result

Indicators, intervention rules, aggregation, thresholds, exclusion criteria, and the population or operating range should be declared in advance. Flexible post-hoc schemes can manufacture almost any classification.

### 5. Allow unresolved outcomes

Sometimes available evidence will not distinguish two plausible accounts of the same episode. “Empirically unresolved” is then more accurate than either a positive or a negative classification. Measurement failure is not automatically evidence that the represented structure was absent.

---

## 6. Candidate research programs

The following studies are possible routes for development, not promises that the project must pursue them all. They differ in maturity and in the claims they would bear on.

### A. Measuring the relationship architecture

**Question.** Can positive and negative inherent value, Relational Positioning, and perceived agency be identified independently? Can access, strict benders, and profile-sensitive contribution be separated from the core? Can the three positioning routes and three identity forms be discriminated?

**Possible design.** Cross positive-only, negative-only, mixed, and neither profiles without collapsing them to a net score. Separately lesion identity-mediated positioning, intimacy, and direct encounter; within identity, lesion shared, contrastive, and complementary-role forms. In further arms, block access while preserving the core, and vary a candidate strict bender while holding the core fixed.

**What would support the application.** Pre-specified indicators and lesions show the predicted route-, form-, and downstream-locus differences on held-out cases and outperform outcome-defined or wholly compensatory rivals.

**What it would not show.** It would not establish a universal threshold, unique indicator set, moral provenance, or the whole theory.

### B. Represented subjects and the third-person subject condition

**Question.** Do robust moral judgment episodes require a represented agent and, when the subject is nonself, does moral subjecthood track represented integrative recognition rather than entity kind?

**Possible design.** Cross human and nonhuman referents with high and low attributed capacities to register the object's standing, represent relevant relations and agency, and integrate these into evaluation. Hold the action, harm, object, and builder profile as constant as possible. Include first-person prospective cases as a control and separately code reconstructive cases that may invite character assessment. A second arm should examine putatively agentless appraisals—including natural-disaster and wild-animal-suffering cases—using pre-specified measures of explicit and implicit subject representation and independent process, transfer, and contrast measures that distinguish moral judgment from tragic or axiological appraisal.

**What would support the application.** Subject-directed appraisal follows the attributed recognition profile more closely than the human/nonhuman label.

**What would create pressure.** A robust episode for which pre-specified measures rule out explicit and implicit subject representation counts against the adequacy of the subject condition when independent process, transfer, and contrast evidence places it with paradigm moral judgments rather than tragic or axiological appraisals. A complete rival strengthens

the challenge but is not required. Entity-kind effects alone may instead reveal a heuristic for attributing the functional subject role.

## C. Moral provenance

**Question.** Can two judgments with the same content and overt response be distinguished by whether a represented moral relationship actually helped form the operative evaluation?

**Possible design.** Compare relationship-constructed judgments, general norms that retain relationship provenance, and free-standing policy or rule application. Combine interventions, transfer tests, process measures, selective disruption and restoration, and held-out model comparison.

**What would support the application.** Relationship-based and policy-based routes show different, replicable signatures while sentence content and overt response are held fixed.

**What it would not show.** A provenance study tests one central claim of the Moral Detachment Thesis. It does not test the whole of NDT, the Constitutive Evaluativity Thesis, replaceable relationship gates, identification of positioning routes and identity forms, frame individuation, challenge admission, or polyphony.

## D. Encoded evaluation and control

Three claims must be tested separately:

1. **Semantic/control-role association:** a representation's semantic performance is systematically associated with independently measured control organization.
2. **Existence of constitutively encoded evaluation:** at least some representations are evaluative partly because of their constitutive role, rather than because a separable evaluator adds evaluation afterward.
3. **Prevalence or non-negligibility:** such constitutively encoded cases occur at a specified rate or above a specified threshold in a declared population or operating range.

Evidence for the first claim alone establishes neither constitutive encodedness nor prevalence.

**Question.** Can semantic/control-role association be identified, and can a constitutively encoded case be distinguished from one explained by a separately tokened desire, policy, or verdict?

**Possible design.** Manipulate semantic discrimination while holding explicit rewards and instructions fixed, and measure a control state independently of the target action or judgment. Transfer to held-out tasks can help distinguish a standing control organization from local performance or associative priming.

**What would support the association claim.** Changes in semantic performance systematically alter independently measured control organization within a declared operating range.

**What further evidence the existence claim needs.** Competing models, interventions, and process evidence must distinguish a constitutive role from evaluation added by a separable evaluator. A prevalence claim then requires a separately justified sampling frame, indicator, and threshold.

**What it would not show.** Semantic success alone does not entail a behaviorally specific ought. Normative detachment still requires alternatives, outcome expectations, standards, feasibility, and guidance.

## E. Normative detachment and violation-sensitive choice

**Question.** When every available option violates something, does represented priority and violation structure guide judgment beyond outcome score alone?

**Possible design.** Hold outcome ranking fixed while changing which requirement is violated and how that requirement is prioritized.

**What would support the application.** The pre-specified fallback model predicts preference for the outcome-inferior option when it violates the lower-priority requirement.

**What it would not show.** Failure of one fallback rule would pressure that implementation family, not the general claim that normative judgment uses represented constraints and guidance.

## F. Frames, plurality, and polyphony

**Question.** Can frame activations and comparison bridges be identified before using them to explain divergent judgments?

**Possible design.** Use changes in object construal, action space, action-guiding question, task regime, or representational similarity to identify frame transitions. Then add or remove represented bridges while holding the local judgments fixed.

**What would support the application.** Pre-specified frame boundaries improve prediction on new episodes, and bridge manipulations move pairs between unaligned plurality and comparison-ready conflict.

**What it would not show.** Several different answers are not automatically polyphony. They may reflect noise, a tie within one model, an ordinary inconsistency, or outputs the person never compares.

### G. Appraisal targets, collective subjects, and strict benders

Several more focused programs are available:

- hold broad valence and event content fixed while varying whether the appraisal targets an act, agent, institution, outcome, relationship, or event;
- hold policy, harm, and member conduct fixed while varying attributed institutional capacities for recognition and revision;
- hold the relationship core and explicit content fixed while varying a candidate downstream factor such as stewardship, guardianship, authority, urgency, vulnerability, or ownership.

These studies could show whether the factor is a strict bender at a registered downstream locus. If it changes inherent valence, positioning, or perceived agency, it belongs instead to core-changing machinery. A label never receives an edge type automatically.

### H. Full-architecture instantiation and provenance-based kind unity

**Question.** Do any real judgment episodes instantiate NDT's complete framed-detachment architecture, and do independently identified provenance groupings form a more stable explanatory kind than lexical, behavioral, character-based, or policy-based partitions?

**Possible design.** Pre-specify independent indicators for frame, evaluative organization, detachment, relationship builders, subject conditions, and actual contribution. Compare a complete-architecture model with component-only and rival partitions across interventions, transfer tasks, process measures, and held-out episodes.

**What would support the bridge claims.** At least some episodes satisfy the complete integration criteria, and provenance partitions yield more stable registered consequences and better out-of-sample comparison than the alternatives.

**What would create pressure.** Sustained failure to identify any complete episode under mature component measures counts against full-architecture instantiation. Repeated failure of pre-specified provenance partitions to yield their registered intervention, transfer, process, or explanatory consequences by itself counts against provenance-based kind unity. Consistent outperformance by rival partitions strengthens that evidence but is not required before it becomes adverse. Mechanistic heterogeneity alone is not decisive because a cognitive kind may be multiply realized.

---

## 7. Moral provenance is important but local

The provenance program is currently NDT's most developed identification proposal. Its central contrast is simple to state:

> Two people may reach the same judgment and give the same answer, while only one judgment was formed through a represented relationship of concern.

No single observation can reveal that difference. A credible study would triangulate several sources of evidence:

- **Intervention:** alter the relationship representation while leaving the explicit rule available, and alter the rule while preserving the relationship route.
- **Transfer:** compare generalization by relational similarity with transfer by membership in a formal rule class.
- **Process:** examine latency, attention, confidence, memory, affective intensity, persistence, and sensitivity to reframing.
- **Selective disruption and restoration:** weaken one route while preserving the other, then restore it.
- **Model comparison:** compare relationship-provenance, policy-only, outcome-only, and mixed-route accounts on held-out cases.

Some third-person and first-person reconstructive judgments may concern the agent's character or the moral worth of conduct rather than the provenance of the focal judgment. That is a legitimate local rival. It should be measured through the appraisal target rather than incorporated into the definition of provenance. First-person prospective cases - in which

someone is deciding what to do before acting - provide an especially useful contrast because they do not reduce naturally to retrospective character assessment.

Even a strong provenance result would remain local. It would support the claim that actual cognitive source can matter to moral type while leaving the rest of the architecture to be evaluated separately. It would not by itself establish full-architecture instantiation or provenance-based kind unity; those are cross-component and cross-episode bridge claims.

---

## 8. What would create pressure for revision?

NDT is answerable to evidence, but the appropriate revision depends on which claim encounters difficulty.

**Claim**: **Encoded evaluation**

**Evidence that would create pressure**: Stable semantic changes have no registered effect on independently measured control organization, or a separately tokened evaluator explains the cases better

**Likely scope of revision**: Restrict or revise the action-guiding or weak empirical versions of CET; this need not decide the strong semantic thesis

---

**Claim**: **Relationship core**

**Evidence that would create pressure**: Independently specified deletion or consequence tests repeatedly show that a supposedly lost distinction remains available without one role; valence components cancel despite the noncancellation rule; or an independently specified rival handles the paradigm and contrast cases better

**Likely scope of revision**: Reconsider the explication, decomposition, or minimality

---

**Claim**: **Represented-subject and third-person subject conditions**

**Evidence that would create pressure**: Pre-specified measures rule out explicit and implicit subject representation in robust episodes that independent process, transfer, and contrast evidence places with paradigm moral judgments rather than tragic or axiological appraisal; or third-person classification systematically tracks entity kind rather than attributed recognition

**Likely scope of revision**: Revise the subject condition, the boundary of the moral kind, or the gate model at the level affected

---

**Claim**: **Provenance**

**Evidence that would create pressure**: Relationship-grounded and policy-only routes remain indistinguishable under the best feasible interventions and models, or policy-only episodes show the predicted moral signatures despite absent relationship contribution

**Likely scope of revision**: Limit the empirical utility, scope, or content of the provenance claim

---

**Claim**: **Full-architecture instantiation**

**Evidence that would create pressure**: Mature component measures repeatedly fit isolated parts, but no real episode exhibits their registered joint organization; or disconnected rival models consistently predict held-out formation and intervention effects better

**Likely scope of revision**: Restrict or reject the full-architecture bridge hypothesis

---

**Claim**: **Provenance-based kind unity**

**Evidence that would create pressure**: Pre-specified provenance partitions repeatedly fail their held-out intervention, transfer, process, or explanatory consequences. Consistent rival outperformance strengthens this adverse evidence but is not required

**Likely scope of revision**: Restrict or reject the provenance-based unifier; heterogeneity alone is insufficient under multiple realization

---

**Claim**: **Fallback**

**Evidence that would create pressure**: Under controlled cases in which all options violate a requirement, judgments track only outcomes despite stable represented priorities

**Likely scope of revision**: Reject or narrow that fallback family, not normative detachment as a whole

---

**Claim**: **Target-sensitive appraisal**

**Evidence that would create pressure**: Target type adds no information beyond valence and wording across pre-specified cases

**Likely scope of revision**: Simplify or revise the appraisal grammar

---

**Claim**: **Alignment and polyphony**

**Evidence that would create pressure**: Adding or removing the relevant comparison bridge has no effect while local outputs remain fixed

**Likely scope of revision**: Revise the proposed bridge structure or its empirical application

---

**Claim**: **Strict benders**

**Evidence that would create pressure**: A candidate factor never changes its pre-specified downstream locus while the core is fixed, or instead changes the core

**Likely scope of revision**: Reject that bender edge or reclassify the factor as core-changing machinery

The four standards locate the consequence: a result may pressure comparative explication, formal countermodel structure, operational identification, or calibrated measurement. Stating the narrowest affected level is part of responsible testing.

---

## 9. What evidence would not show

Several tempting inferences are too strong.

- A person's use of the word *moral* does not by itself identify the cognitive type of the judgment.
- Moral-looking behavior does not show that a moral judgment produced it.
- A rule, utility calculation, sacred value, or philosophical label does not establish moral provenance.
- A self-report of care, route uptake, agency, or motive is evidence, but not a sufficient independent measure.
- A study compatible with NDT is not necessarily a test of NDT.
- Successful execution of formal test cases shows that a finite model conforms to registered specifications; it is not evidence that human cognition has the same structure.
- Failure to measure a latent role does not by itself show that the role is absent.
- Evidence for one part of the theory does not confirm the entire architecture.
- Failure of one quantitative threshold or provisional algorithm does not refute the constitutive distinction it was meant to realize.
- Conceptual coherence alone does not establish prevalence, calibrated measurement, or implementation in a particular population.

These limits apply in both directions. They prevent overclaiming in favor of NDT and prevent local or measurement-level failures from being inflated into conclusions they do not support.

## 10. A recommended order of development

There is no requirement that NDT become a single laboratory program, and empirical completion is not the master criterion of its value. If empirical work is undertaken, however, several dependencies suggest a sensible order.

1. Develop independent measures of semantic control recruitment, rather than defining it by the target behavior.
2. Establish independent indicators for both inherent-valence components, Relational Positioning, and perceived agency before fitting any population-specific gate.
3. Use validated relationship-role indicators before testing the third-person subject gate.
4. Extend the provenance program through intervention, transfer, selective disruption and restoration, and held-out model comparison.
5. Develop independent frame-individuation measures before making population-level claims about polyphony.
6. Treat fallback, target-sensitive appraisal, collective subjecthood, and strict-bender edges as separate local programs.
7. Test full-architecture instantiation and provenance-based kind unity only after component measures and local provenance models are sufficiently mature.

This is a dependency map, not a commitment to conduct every study. Some questions may remain more productive as conceptual or comparative inquiries. Others may become tractable as methods improve.

## 11. The present evidential position

NDT's current strengths are conceptual, architectural, and formal. It takes a classification problem as its starting point, proposes a formally and conceptually noncircular architecture for distinguishing moral from nonmoral normativity, separates judgment content from the source that formed a particular token, and states consequences that alternative accounts can contest. Whether that boundary can be identified through validated, operationally independent measures remains open. The companion Moral Cognition Explanandum Audit assesses how a representative group of existing programs handles that classification problem; it does not claim exhaustive coverage of every possible theory. The executable model demonstrates finite conformance for selected cases and boundaries. None of this is direct empirical confirmation.

### The executed study: a bounded result

The executed study (Beal & Rottman, 2024) comprised two preregistered studies of U.S. adults (combined final $N = 361$) comparing issue-targeted nonmoral-belief ratings with broad moral-value measures across eight controversial judgments. In the authors' reported regressions, the targeted belief items outperformed the study's broad value measures for each judgment. Study 2 also found weaker associations for generic "attractor" beliefs and repeated the comparison using the Moral Foundations Questionnaire. The narrow supported result is that broad value endorsements left issue-level variation unexplained while beliefs about the relevant situations carried substantial additional information.

The design does not establish why the targeted items performed better. Each bender was written for a particular judgment and predicted direction, whereas broad value items were reused across judgments. Greater target specificity and semantic proximity could therefore explain part of the advantage. A validation vignette showed that participants did not treat a bender sentence as simply synonymous with its associated moral verdict: they expected the verdict to change when another relevant premise changed while the bender was held fixed. That result does not establish psychometric discriminant validity, exclude a common latent attitude or shared method variance, or show that the rated belief actually contributed to the judgment episode. Study 2's generic attractors weaken a wholly specificity-based explanation but do not eliminate these alternatives.

Item construction and analysis add further limits. The benders were tailored to particular outcomes, while the broad values were general; Study 2 then reworded low-performing items in both classes, making it a tuned extension rather than a fully frozen replication. Both studies used backward stepwise regression, eliminating predictors by $p$-value until the retained predictors were significant; Study 2 then recombined selected predictors in further models. Preregistration and Bonferroni correction address some risks, but they do not provide post-selection inference, model-stability evidence, or held-out prediction. Counts of retained benders and values are not exchangeable opportunities when the item families differ in number, construction, and specificity.

The studies were cross-sectional, and participants made their moral judgments before rating the candidate beliefs. Their associations therefore cannot distinguish belief-to-judgment influence from post-judgment rationalization, consistency maintenance, reciprocal influence, or a common cause. Exploratory controls for political conservatism narrow one rival explanation but do not settle causal direction.

For NDT, the study supplies bounded evidence consistent with distinguishing candidate benders from broader attractors and motivates the registered robustness-ordering question. It does not test the current relationship core, the subject condition, actual contribution, the Moral Detachment Thesis, the matched-perturbation ordering, full-architecture instantiation, provenance-based kind unity, or NDT as a whole. A stronger follow-up would develop items in a separate sample; cross moral or nonmoral status with target specificity; freeze the comparison model; evaluate held-out prediction; and manipulate or temporally separate candidate beliefs and judgments.

The project has not yet demonstrated that a real judgment episode instantiates the complete architecture or that provenance partitions form a superior empirical kind.

Empirical risk is therefore distributed. The provenance-based kind-unity hypothesis bears the most concentrated risk for NDT's proposed unification; the whole-architecture hypothesis bears the corresponding risk for full instantiation. The comparative-explication proposal is assessed through comparative explication and independently specified deletion or consequence tests, while the local empirical hypotheses face the intervention, transfer, process, and measurement evidence registered for their own domains.

Operational identification is an open area of development. Measurement is least complete for independent inherent-valence components, positioning routes and identity forms, functional access, strict-bender edges, profile-sensitive contribution, replaceable application gates, and the subject threshold. Actual contribution, implicit subject representation, frame boundaries, and challenge admission are latent and may sometimes remain difficult to identify.

Falsifiability matters where NDT advances an empirical prediction, a quantitative model, or a replaceable implementation. It is not the sole or central standard of the theory's value. Clarifying the explanandum, proposing a formally and conceptually noncircular architecture, preserving important distinctions, and exposing dependencies that other theories leave hidden are substantive aims. Empirical inquiry can strengthen, refine, restrict, or undermine applications of that architecture. It need not pretend that every constitutive or semantic question can be settled by one experiment. Where NDT makes descriptive bridge claims, sustained failure on their registered consequences can require restriction or abandonment.

---

## Appendix: Technical crosswalk

Readers who want the exact claim classifications can consult **Formal Specification 1.6.3**, **Proposition Registry 2.9**, and **Empirical Audit 2.3**. The current comparison companions are **NDT in Context Dataset 1.20**, **Data Contract 1.13**, and **Literature Review 0.13**; **Executable Model 2.2** is a finite, non-authoritative companion. The most relevant registered topics are:

- the possibility of encoded evaluativity, together with separate empirical commitments concerning its existence and non-negligible occurrence;
- constitutive relationship-role explication;
- empirical versions of encoded evaluation and its prevalence;
- measured application of the relationship model;
- measured applications of the represented-subject condition and the third-person gate;
- moral provenance, actual contribution, empirical unresolvedness, and cross-scheme comparison;
- local hypotheses concerning challenge admission, frames, polyphony, collective subjecthood, relation qualifiers, fallback, appraisal targets, alignment, and robustness;
- full-architecture instantiation;
- provenance-based kind unity;
- independent identification of control recruitment;
- the broad, distributed empirical program;
- comparative explication adequacy;
- the three positioning routes, three identity-mediated forms, and their distinguishing transfer predictions.

The formal architecture controls the definitions. The entries above distinguish claims about the modeled kind from hypotheses about real systems and from proposals for identifying latent structure.