

Reviewing Normative Detachment Theory

Bree Beal

A guide to the questions, sources, and standards of criticism

Version 1.12 · September 2026 · Author draft for discussion

Normative Detachment Theory proposes an account of how judgments form and what makes some normative judgments moral. Its central question is not whether an action is right, whether a motive is admirable, or whether a person uses moral words. It concerns the organization of a particular occurrence of judgment: how someone understands a situation, what matters within that understanding, and what actually helps form the conclusion.

A useful review can challenge the proposed organization, the argument for it, the formal consequences, the evidence, or the explanation offered to readers. Accurate reconstruction is necessary for a criticism to reach its target. Agreement with the theory is not.

Human readers are welcome to ask questions, report confusion, or raise objections at any point. No one needs to master the formal specification to show that an introductory paragraph is unclear. The more demanding preparation requirements below apply to AI systems commissioned to review or revise the theory. They are safeguards against mistaking a familiar substitute for the account actually proposed.

1. Choose a route through the project

For an introduction, start with the [Reader-Facing Article](#). It develops a small set of examples to explain judgment formation, the structure of a moral relationship, and conflicting interpretations. The [Normative Cognition Explorer](#) provides interactive demonstrations of the same distinctions. Neither replaces the formal sources when an exact condition is at issue.

The [Formal Specification](#) states the definitions and dependencies. [Scope and Commitments](#) explains their intended domain. Reading both in full is required for commissioned AI architecture reviews and recommended for human readers investigating the complete architecture; a local reading can miss a condition supplied elsewhere. Human participation does not depend on completing that reading. The [Proposition Registry](#) distinguishes assumptions, proposed definitions, formal consequences, empirical hypotheses, and open identification problems.

For evidence, begin with [Evidence and Open Questions](#). It explains the completed study, the inference it supports, and the difference between an explicit hypothetical route and a route identified in a real person. The [Empirical Audit](#) supplies the fuller technical catalogue. A proposed research design in that catalogue is not evidence that the required measures have been validated.

For comparison with other work, use the [Literature Review](#) and [Bibliography](#). The review develops intellectual encounters and preserves a record-level reference section. The underlying comparison data, source register, and data contract distinguish exact source passages from the claims NDT makes about them. Consult those records and the original source when challenging a particular mapping. A citation proves neither equivalence nor influence merely by appearing in a bibliography.

The [Moral Cognition Explanandum Audit](#) addresses the question of what selected research programs count as moral cognition. It supports a bounded diagnosis of differing explanatory targets, not an exhaustive claim about everything ever written in moral psychology.

2. Locate the object of disagreement

Several questions are close enough to be confused but different enough to require separate arguments.

The judgment and the action

A parent may judge that spoiled milk should be discarded and then be interrupted before discarding it. Someone else may discard milk as part of an automatic routine without forming the judgment under review. NDT concerns judgment formation; action has additional conditions. An objection based on behavior should explain the bridge from what the person did to the thought the theory claims to describe.

This does not make behavior irrelevant. It makes the inference from behavior explicit. Equally, the difference between thinking and outward performance is not established by a single successful experiment. It is part of identifying what a cognitive explanation is trying to explain.

The moral classification and the value of the judgment

NDT uses *normative* broadly for prescriptions and appraisals. A judgment about efficient tool use or a sensible routine can be normative without being moral. Calling a judgment nonmoral does not call it bad, selfish, deficient, or inappropriate. Calling a judgment moral does not endorse it: cruel or mistaken judgments may be morally organized in the descriptive sense at issue.

A criticism may challenge that proposed boundary. It should not first replace the boundary with moral approval and then fault NDT for the consequences. Moral worth, responsibility, and character are legitimate topics; they are not interchangeable with the cognitive type of the focal judgment.

The conclusion and the contributing thought

NDT allows two judgments with the same content to differ in what helped form them. In the breakfast example, concern for the child may help make serving the milk unacceptable, while a separately stipulated procedural judgment may rely entirely on a food-handling rule. The parent can care in both episodes. The difference concerns the use of that relationship in the particular judgment.

Rule-following alone does not distinguish the two routes. A rule may express or preserve relationship-based concern. Nor does moral cognition require a pure motive: moral and nonmoral considerations can both contribute. The target is contributing organization, not a test of whether the person had only admirable reasons.

The formal account is deliberately stricter than finding a relationship somewhere in a person's thoughts. The relevant relationship and the exact subject and object must help organize the same judgment-forming evaluation. A relationship that is merely available, or contributes to an unrelated thought, cannot supply that condition. Several routes may contribute redundantly even when removing one would leave the spoken conclusion unchanged.

The proposed organization and its measurement

Defining a role does not supply a validated instrument for measuring it. NDT's relationship account distinguishes inherent value, relational positioning, and perceived agency. It does not assign one universal questionnaire, numerical threshold, or brain signature to each role.

A critic can challenge the role itself or a proposed way of identifying it. Those challenges have different consequences. An inadequate measure does not by itself show that the role is unnecessary; an intelligible distinction does not establish that its occurrence in a person can currently be identified with precision. The present project supplies no validated detector for all the exact internal contribution conditions of ordinary human judgment.

3. Examining the relationship account

The relationship structure carries much of NDT's explanatory burden. It deserves reconstruction before either acceptance or rejection.

An object can matter to someone in its own right, positively or negatively, rather than only as a useful means or an obstacle. That is the question of inherent value. The positive and negative components remain distinct: seeing something as both good and bad is not the same as seeing it as neither. NDT does not collapse the pair into one net score.

A separate question concerns how the object stands in relation to the represented subject. NDT calls that organization *relational positioning*. The account distinguishes identity-mediated positioning, intimacy, and direct encounter. Identity-mediated positioning includes shared identities, contrastive identities, and complementary roles. A physician and patient can be connected through an operative role relation without being intimate or sharing a group identity. Merely assigning the words “physician” and “patient” does not establish that the connection is used in the relevant thought.

Perceived agency concerns the subject's represented capacity to affect or respond to the object. That is not identical to whether one specific action is feasible at the moment, nor to whether the subject deserves blame. A response can be prospective, reparative, expressive, or otherwise situated within the possibilities the person represents. The relevant range cannot be expanded after the fact simply to save a preferred classification.

The theory treats these as three builders of the relationship. Derived relationship strength is not a fourth builder. When another agent is represented as the moral subject of the judgment, a further condition concerns the capacity attributed to that agent to recognize and integrate the relevant considerations. Entity kind alone—human, animal, institution, artificial system—does not settle that condition. Self-directed cases still need the ordinary relationship and contribution conditions; self-reference does not turn all prudent self-management into moral judgment.

Removal arguments ask whether each role does distinguishable work. Removing every qualifying positioning route is different from removing intimacy alone: an operative identity or encounter may still connect the subject and object. Blocking a route's access to the judging process is different from removing the relationship it would have carried. A critic should use the comparison appropriate to the claim rather than change several conditions and assign the result to one of them.

Changes in what follows from a relationship raise a different question. NDT calls an operation a *bender* when it changes a downstream requirement, priority, appraisal, or other consequence while preserving the exact relationship core and its continuing contribution. An operation that changes the relationship core is not bending in that sense. The same representation can perform different roles in different operations; a label such as ownership, urgency, or vulnerability does not settle the role.

These are proposed explanatory distinctions, not a demand that every rival adopt NDT's vocabulary. A rival can redistribute the work. It can also argue that a distinction is unnecessary. The substantive comparison asks what each account can represent and explain under matched conditions.

4. Different criticisms require different support

An internal formal criticism asks whether a consequence follows under the stated definitions. A conceptual criticism may challenge those definitions, their independence, their scope, or the explanatory purpose they serve. A proposed counterexample should make clear which kind of argument it is. Showing that a definition produces an unfamiliar classification can begin a conceptual challenge; unfamiliarity alone does not complete it.

An empirical criticism should identify what was observed, how it bears on the proposed organization, and which alternatives remain. Behavioral fit does not reconstruct an internal route, and failure to detect a latent representation does not establish its absence without an adequate detection model. Conversely, a complete superior theory is not required before adverse evidence matters. An unexpected result can be worth investigating even when it was not predicted or preregistered. The inference should be proportionate to the evidence available.

A literature criticism should reconstruct the relevant external account at the appropriate developmental stage. An early instrument is not necessarily the mature theory. A precise comparison with an early paper can still misrepresent a lineage if later qualifications change the target. Compare a definition with a definition, a mechanism with the relevant operation, or a measurement with a matched measurement; explain any deliberate move between levels. Shared vocabulary, missing vocabulary, or a shared example does not decide the relation.

A communication criticism has its own legitimate target. If a reader cannot tell whose perspective a paragraph describes, what an example assumes, or how a conclusion follows, the difficulty is not answered merely by pointing to a distant formal definition. The source may settle the intended commitment while leaving an exposition problem to repair.

5. Preparing a commissioned AI review

An AI commissioned to assess the architecture or revise its theoretical exposition must first read the complete designated Formal Specification and Scope and Commitments. It must also read the additional authorities needed for its assignment. Historical papers, summaries, search hits, and recollection are not substitutes for the current account.

The preparation record should identify document versions and actual reading coverage. A complete, verified text extraction is acceptable; creating an extraction is not the same as reading it. A verified

comparison can update a fully read prior version only when the prior coverage is known and every intervening change has been checked in context. Missing or unread material should be disclosed.

After reading, the reviewer should give a compact reconstruction of the load-bearing distinctions and the specific target of review. The reconstruction should apply the distinctions to a case, not merely repeat their names. Questions, ambiguities, and apparent contradictions should be recorded during preparation rather than suppressed. The purpose of the checkpoint is to establish the target accurately, not to obtain agreement with the theory or approval of the coming criticism.

Once the reconstruction meets the agreed preparation requirements, substantive review proceeds. A reviewer may accurately reconstruct a claim and then argue that it is false, circular, unnecessary, unidentifiable, or inferior to another account. A source passage that resolves an initial misunderstanding does not automatically dispose of the objection to the corrected claim.

Human readers need not follow that commissioned-AI procedure to participate. Before a broad formal or public critique, a source-prepared discussion with a capable frontier AI can be a useful way to check for misconstrual. Direct source reading or consultation with another person is equally legitimate. An AI's agreement is neither required nor a guarantee of correctness. The human critic retains responsibility for the argument.

6. Make the criticism easy to evaluate

A report should identify the claim in ordinary prose, give an exact source locator, explain the objection, and state its warranted consequence. Stable record identifiers can support traceability, but should not replace the explanation. “The agency condition excludes a case the Article appears to include” tells a reader what to investigate; a bare code does not.

For a worked example, a useful report states which facts are assumed, which feature changes, which conditions remain, and why the conclusion follows. For a source comparison, it states the external author's claim, the NDT claim, their respective levels, and the proposed relation. For empirical evidence, it distinguishes the observed result from the inferred process. For a formal defect, it supplies the conflicting clauses or countermodel.

The consequence should be neither artificially narrow nor inflated. An ambiguous sentence may need only a clearer explanation. A missing formal condition may affect several theorems and illustrations. A new comparison may require source verification without requiring theory revision. Explain the dependencies that determine the scope of the correction.

Recommendations remain recommendations until authorized for implementation. A review, an editorial correction, a passing executable test, and publication of a discussion draft are not interchangeable with ratification of the theory. AI-assisted artifact-integrity review is also not a substitute for independent specialist or human peer review.

7. A final check before submitting

A substantive report should make the following questions answerable:

- What exact version and claim did the review assess?

- Does the report distinguish the judgment from its expression, action, moral worth, and correctness?
- Does it reconstruct the proposed cognitive route without inferring that route from its conclusion alone?
- Is the objection conceptual, formal, empirical, comparative, or communicative, and what support is appropriate to that target?
- If another theory is invoked, does the comparison preserve its developed commitments and the relevant explanatory level?
- If a removal or intervention is used, are the changed and preserved conditions explicit?
- Are uncertainty, uninspected sources, and unresolved interpretations visible where they affect the conclusion?
- Does the proposed correction preserve the theory's other commitments, or does it explicitly request a substantive decision?

NDT invites criticism of its central argument, not only small repairs at the edges. A review succeeds when it makes the pressure point clear enough that readers can distinguish a misunderstanding, a genuine disagreement, and a demonstrated defect—and can see why the distinction matters.