

Evidence and Open Questions

Bree Beal

What the argument establishes, what has been studied, and what remains unknown

Version 1.12 · September 2026 · Author draft for discussion

Two people can agree that an action is wrong while understanding the situation very differently. Conversely, people who both endorse care or fairness can disagree about what a particular situation calls for. Normative Detachment Theory asks what a person represents, how those representations fit together, and which of them help form a judgment. Explaining that organization is a different task from predicting a spoken answer.

The distinction matters when assessing the theory. An imagined case can make two ways of thinking intelligible without measuring either one. A formal example can show that the proposed architecture preserves a distinction without establishing how often people instantiate it. A survey can reveal an association between beliefs and judgments without recovering the thought that produced a particular response. Each achievement can be useful; none should be made to stand for the others.

NDT's present case is primarily conceptual and architectural. Its formal companion makes dependencies explicit and implements selected cases. A completed two-study project also bears on one important question: how much variation in controversial judgments is associated with beliefs about the situations being judged, compared with broad endorsements of moral values? The studies do not establish NDT's complete architecture or identify the exact cognitive route of an individual judgment.

This document begins with that completed research, then explains the different kinds of work still needed. The [Reader-Facing Article](#) develops the argument itself. The [Empirical Audit](#) records the more technical hypotheses and identification proposals.

1. Why beliefs about a situation deserve attention

Suppose two people both believe that protecting people from avoidable harm is important. They are considering a public policy. One believes the policy will protect vulnerable people; the other believes it will expose those people to danger. Their shared concern does not settle their disagreement, because they disagree about what the policy does.

That ordinary possibility motivates a research question. When people disagree about an issue, how informative is a general statement of their values, and how informative are their beliefs about the relevant people, actions, and consequences? The question does not presume that values are irrelevant. It asks what remains unexplained when a study measures values without adequately measuring the situation as participants understand it.

Beal and Rottman's submitted manuscript, *Nonmoral beliefs are better predictors of moral judgments than moral values* (2024), examines that comparison across eight controversial issues. The project-held

manuscript is [available with the earlier NDT research](#). The account below reports its analyses rather than a new analysis of the data.

What participants were asked

The first study recruited 200 adult United States residents through Prolific in May 2023. After seven exclusions for attention checks or missing data, the analysis used 193 participants. The second recruited 200 adults in July 2023 and retained 168 after attention-check exclusions. The combined final sample was 361. The manuscript reports preregistration for both studies.

Participants first rated judgments about abortion, police defunding, environmental protection, vaccine mandates, gun rights, the travel policy described in the questionnaire as Trump's "Muslim Ban," cancel culture, and transgender bathroom restrictions. They then rated candidate explanatory beliefs and values. The order of the later measures and the items within them was randomized; the moral judgments nevertheless preceded the proposed explanatory ratings.

The comparison deliberately placed different kinds of statement beside one another. A broad value item asked whether preventing harm was of utmost moral importance. An issue-specific item asked, for example, whether police forces keep neighborhoods safe. The first concerns an endorsed value across situations; the second concerns how a particular institution affects the world. Participants could agree about harm while disagreeing about the police, and the study asked which ratings were more strongly associated with their policy judgments.

For each issue, Study 1 included two targeted belief items, six broad value items, and ratings of the moral worth assigned to two relevant entities. Study 2 omitted the moral-circle measure, added ten more general beliefs associated with ideological outlook, and included the original Moral Foundations Questionnaire as well as the authors' own broad-value items. Study 2 also reworded four low-performing belief items and four low-performing value items from Study 1. It was therefore a related, revised comparison, not an exact repetition with every item frozen.

What the questionnaires measured

Order of questions—not an observed sequence of thinking



Study 1 · 193 participants

Targeted beliefs, broad values,
and moral-circle ratings

Study 2 · 168 participants

Revised items; foundations questionnaire
and general ideological beliefs added

The analyses compare reported associations.

They do not identify the internal route of any individual judgment.

Figure 1. The sequence is the questionnaire sequence, not a discovered sequence of thought. Participants rated the judgments before the beliefs and values used to explain them statistically. Study 2 added the Moral Foundations Questionnaire and general ideological beliefs. No part of the design selectively removed an internal reasoning route.

What the analyses found

For each issue, the authors fitted a regression relating judgment ratings to the candidate predictors. They used backward stepwise selection: predictors were removed according to their statistical significance until the remaining set met the selection criterion. The reported tables use a Bonferroni-corrected significance threshold of .00625 for the eight issue comparisons.

In Study 1, at least one targeted belief remained significant for every issue. The authors call a targeted belief the best predictor on each of the eight issues, explicitly defining “best” by the highest *t*-value and lowest *p*-value in the reported comparisons. That is a comparison of in-sample statistical associations. It is not a demonstration that the beliefs predicted new participants, new issues, or future judgments more accurately.

Study 2 again reported an advantage for targeted beliefs on every issue, both against the authors' broad-value questions and when Moral Foundations Questionnaire scores replaced those questions. The more general ideological beliefs also showed associations with judgments, although they performed less strongly than the targeted beliefs in the reported combined models. Additional analyses controlling for social and economic conservatism were exploratory rather than preregistered. In those analyses, targeted beliefs generally retained more associations than the other measures.

The narrow result is important enough without a stronger gloss: in these samples and analyses, knowing participants' issue-specific beliefs helped distinguish their judgments in ways the measured broad values

did not. The result gives researchers a reason to examine people's understandings of situations closely. It does not demonstrate that values do no cognitive work, or that the measured beliefs caused the judgments.

Why the advantage needs interpretation

The targeted beliefs were written for particular issues and expected directions of judgment. The same broad values were reused across issues. Greater specificity and closer meaning therefore provide a live alternative explanation for some of the difference. If a questionnaire asks a broad question on one side and a carefully matched question on the other, their relative performance cannot by itself isolate the feature that makes one item family theoretically distinctive.

The authors addressed one part of that problem through a validation vignette. Participants imagined someone who retained a focal belief while changing another relevant premise, and they expected the person's judgment to change. A statement such as a belief about when life begins was therefore not treated as simply interchangeable with the associated abortion judgment. That result helps distinguish a premise from its putative conclusion. It does not establish that the two questionnaire measures are psychometrically independent, rule out a common underlying attitude, or show that the belief actually contributed to any participant's own judgment.

The more general ideological items in Study 2 complicate a wholly specificity-based explanation: some beliefs not tailored to a single issue also carried information about judgments. They do not remove the difference between the construction of the targeted beliefs and the broad-value measures. Nor do statistical controls settle the direction of influence. Because the design was cross-sectional and judgments came first, the ratings remain compatible with several possibilities: beliefs influencing judgments, judgments shaping later reports of belief, mutual influence, a wish to give consistent answers, or an unmeasured factor affecting both.

Model selection creates another limit. Selecting predictors and assessing them in the same sample can make the retained model look more stable than it would be in new data. Preregistration and correction for multiple comparisons address important risks, but neither supplies held-out prediction or a complete account of uncertainty after selection. Counts of retained belief and value items are also not a fair substitute for a matched comparison when the item families differ in number, specificity, and construction.

Finally, the sample and issues matter. These were online studies of United States adults responding to politically charged statements. The results do not establish population-invariant effects, cross-cultural measurement equivalence, or the prevalence of any particular kind of moral reasoning. The design does not isolate possible effects of issue wording, political cues, demand, or shared response method. Those are limits on the inference, not reasons to ignore the associations that were observed.

What the study contributes to NDT

The studies support taking situation-specific beliefs seriously when explaining disagreement. They are consistent with the proposal that what someone believes about a situation can matter beyond general value endorsement. They do not establish the stronger proposition that the measured beliefs are the actual premises of the relevant judgment episodes.

The manuscript calls its targeted beliefs *benders*. In NDT's exact structural use, however, a bender changes what follows from a relationship while the relationship core remains fixed and continues contributing. The studies did not establish those preservation and contribution conditions. An item could instead change how a person understands the object or relationship, express an existing attitude, or have several roles. The historical label does not decide which operation occurred.

For the same reason, the study is not a test of the three relationship builders, the requirement for a represented moral subject, or NDT's moral/nonmoral boundary. It also does not test the proposed robustness ordering of builders, benders, and influences on frame selection: the perturbations and measurements needed for that comparison were not matched. A result can motivate a structural account without already verifying it.

Source details: submitted manuscript, numbered pp. 8–10 for Study 1 design and analysis; pp. 14–16 for Study 2; pp. 19–25 for combined and exploratory analyses; pp. 7–8 for the validation vignette; pp. 26–27 for the authors' discussion and limitations. The supplied PDF includes a cover, so its viewer page number is one greater. The preregistration and data records are linked in the manuscript; the present document does not report an independent audit of those records or a statistical replication.

2. An argument about cognition is not a result about behavior

Consider a parent preparing a child's breakfast. The parent notices that the milk smells sour, believes that drinking it could make the child ill, and judges that the milk should be discarded. We can imagine the parent's concern for the child helping make serving the milk unacceptable. We can also imagine a judgment formed solely by applying a familiar food-handling rule. The parent need not cease caring about the child for a routine to settle a particular practical question without that concern contributing.

The imagined difference is not visible in the discarded milk. Both judgments can lead to the same sensible action. Nor does the procedural judgment become a bad judgment merely because NDT would not classify its stipulated organization as moral. A procedure may handle the situation perfectly adequately.

This contrast illustrates why an explanation of thinking cannot always stop at what someone says or does. A person can reach an answer by recalling it, calculating it, copying it, or reasoning from considerations that make it intelligible. Describing the outward answer leaves those possibilities open. Reflection on such differences can clarify the object of a cognitive explanation without requiring an experiment to establish that thinking and outward performance are different things.

NDT makes a further proposal. Among judgments with the same conclusion, the contribution of a particular represented relationship can matter to moral classification. That proposal does not follow merely from the general distinction between thought and behavior. The theory must explain why its relational organization is the appropriate boundary, what distinctions the boundary preserves, and why alternatives are less adequate for that purpose. Those are central tasks of the Article and Literature Review.

An imagined case helps by making the proposed reasoning explicit. It does not give an investigator privileged access to a real parent's cognition. Introspection can reveal differences a person recognizes in thinking, but a report about a reason is not automatically a complete account of every contributing

process. These limits should shape research expectations without displacing the conceptual argument that makes the research question intelligible.

3. Four questions, four kinds of answer

Is the proposed distinction worth making?

NDT proposes to explain moral cognition as a more specifically organized part of normative cognition. Its adequacy depends on whether the proposed boundary is clear and noncircular, whether it handles recognizable cases and contrasts, and whether its parts explain something together that their labels alone do not explain.

For example, treating an object as good in itself differs from treating it as useful. Representing oneself as able to respond differs from passively appreciating the object. Understanding how the object stands in relation to oneself differs from simply having both oneself and the object in mind. NDT makes those distinctions separate and then asks how their combination helps form a judgment. The reader can examine the distinctions, challenge their necessity, and compare alternative organizations.

A definition is not vindicated merely because examples can be described in its vocabulary. A serious challenge may show that the supposed distinctions are unnecessary, that a counterexample exposes a missing premise, or that a rival preserves the explanatory work with a different structure. Equally, an ordinary use of the word *moral* does not settle the very boundary being debated. The argument must engage what the case contains, not just count labels.

Does the formal account preserve the distinction it states?

Formalization makes dependencies inspectable. If the theory requires a relationship to help form a judgment, a formal example must not satisfy that requirement merely because a relationship appears somewhere in the represented situation. The relevant subject, object, and relationship must actually contribute together in the specified way.

The executable companion checks selected finite examples and deliberate counterexamples. Such tests can expose a clause that accepts the wrong combination of conditions or fails to preserve a stated distinction. Passing them establishes behavior of the tested implementation. It does not prove the entire quantified theory, certify every correspondence between prose and code, or show that a human episode has been reconstructed correctly.

The same limit applies to the Explorer's removal demonstrations. When all qualifying connections between a represented subject and object are removed, the illustration shows what becomes unavailable under the stated architecture. The demonstration is a controlled change in an explicit example. It is not a claim that a laboratory can selectively remove the corresponding thought from a person while holding every other relevant process fixed.

Can the organization be identified in a real episode?

An investigator has observations, not the stipulated internal description of a thought experiment. To distinguish a relationship that contributed from one merely available in memory, evidence would have to

discriminate those alternatives. A verbal explanation, reaction time, confidence rating, attention measure, or brain signal does not do so just by being collected.

The project does not supply a validated method with the fidelity needed to identify all of NDT's exact contribution and binding conditions in ordinary human judgment. Suggestions to combine interventions, process measures, and transfer tasks are research proposals, not available instruments. No exact-route verdict is warranted simply because several imperfect measures agree if the same untested assumption links each measure to the proposed route.

This is an identification limit. It neither proves that the stipulated cognitive difference is unreal nor confirms NDT's particular interpretation. An honest account can remain unresolved about a person's judgment while being precise about the alternatives at issue.

Can identification support reliable quantities and generalizations?

Even an identified instance would not establish prevalence, a universal threshold, or a quantitative interaction. Measuring positive and negative inherent value separately, for example, would require evidence that each measure captures the intended relation to the object rather than general liking, distress, approval, or a desired judgment. A useful measure would also need a declared population and setting in which its interpretation remains justified.

NDT fixes no universal numerical cutoff for inherent value, agency, or relationship strength. The separate positive and negative components cannot be replaced by one net score, because both-positive-and-negative is not the same state as neither. Choosing a particular scale or threshold is additional work, not a number hidden in the definition waiting to be read off.

Different questions need different kinds of support

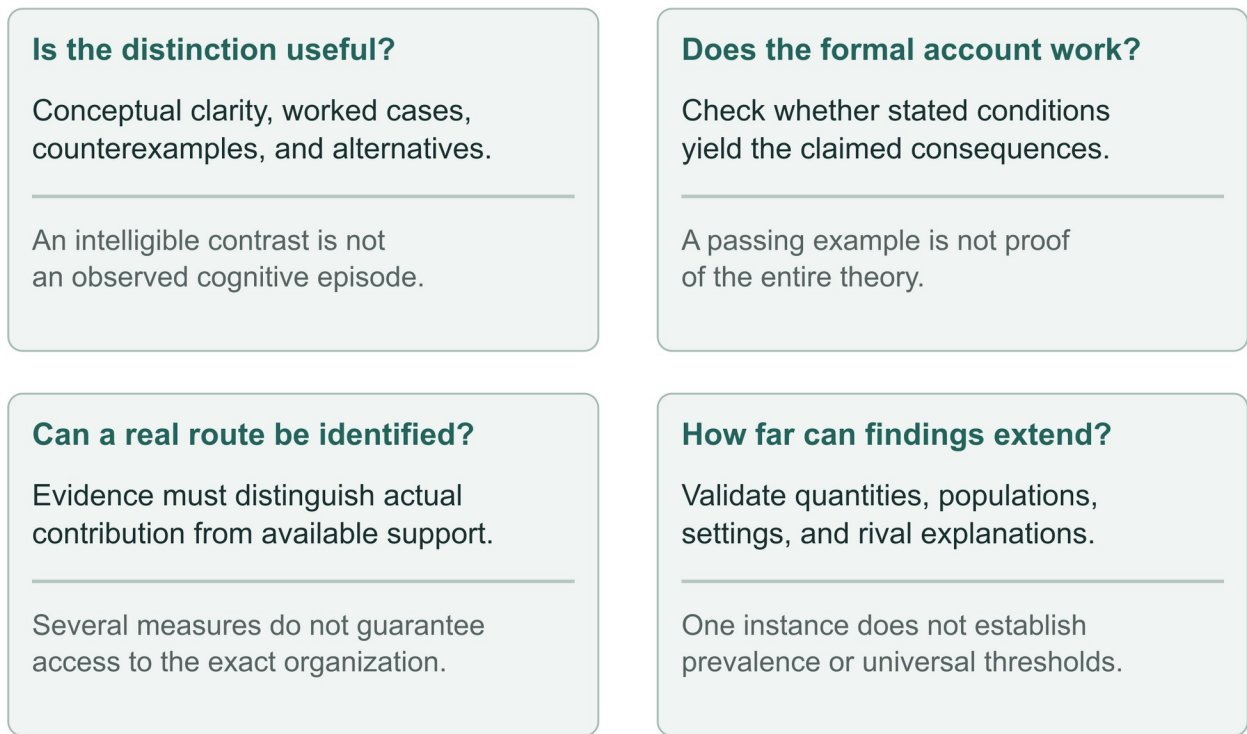


Figure 2. The four questions are not a ladder on which every inquiry must begin at the bottom. They distinguish what an answer establishes. A formal construction does not become an observed cognitive episode, and an observed association does not become an identified reasoning route, merely by being placed beside the theory.

4. Productive next work without imaginary instruments

Several useful projects do not depend on complete access to internal cognitive routes. Others require methodological advances that should be stated before promising the test.

A stronger comparison of beliefs and values

The completed study suggests a tractable follow-up. Develop the items in a separate sample; match candidate belief and value items more closely for specificity; freeze the analysis before testing new participants or issues; and compare predictive performance on held-out data. Such a design could ask whether the observed advantage survives more balanced measurement and genuine generalization.

A further design could vary information about a policy while leaving its stated objective unchanged. In a fully introduced scenario, participants might learn different facts about whether a policy achieves the protection it promises. If judgments change, the result would provide evidence about the effect of that information under those conditions. It would not establish that the information acted only as a bender. The information could also alter trust, perceived agency, the represented relationship, or the participant's interpretation of what response the experimenter expects. Measurement and rival explanations would determine how far the result could go.

The value of this work is not that it settles every theoretical question. It tests a specific explanatory comparison that can be made more rigorous using available survey and experimental methods.

Clearer conceptual and comparative tests

The relationship account can be examined through carefully specified cases. Does removing inherent value leave mere usefulness? Can the same object be treated as both inherently good and inherently bad without either evaluation disappearing? Does an identity connect the relevant subject and object, or is a category label being mistaken for that connection? What changes when a person can appreciate an object but represents no way to respond to it?

These questions should be worked through rather than answered with a list of theory terms. A rival need not use NDT's names. It may distribute the same work differently or reject a distinction with reasons. A useful comparison identifies exactly what each account can express, which assumptions it needs, and what follows from changing a shared example. Formal counterexamples and conceptual disagreements can be informative even when their corresponding human states cannot yet be selectively manipulated.

The literature also supplies more local empirical questions. Measures of intimacy, group identification, harm perception, or attributed capacity may illuminate parts of a proposed relationship. They are not automatically measures of NDT's exact constructs. The [Literature Review](#) distinguishes those encounters from whole-theory confirmation.

Limited tests of guidance, conflict, and appraisal

A fully explained choice can separate expected outcomes from represented requirements. Suppose someone believes that either available action will violate a commitment, but regards one commitment as having priority. A study could test a specified account of how that priority affects the chosen recommendation while holding the described outcomes constant. Failure would bear first on that account of priority and fallback, not on every part of NDT.

Likewise, an appraisal of a harmful event need not be an appraisal of an agent's blameworthiness. A study can vary information about knowledge, control, or responsibility and ask whether participants distinguish the event from the person. Such work can investigate a relevant distinction without claiming to have detected the absence of every implicit subject representation.

Framing and polyphony require comparable restraint. Two answers are not automatically a conflict: they may concern different actions or questions. A scenario can make two interpretations explicit and ask whether participants recognize a common act and incompatible answers to the same question. That task can examine reported understanding. It does not by itself identify the person's entire underlying frame organization or demonstrate that the two judgments were simultaneously active.

Exact provenance identification remains conditional work

The formal account records hypotheses about discriminating relationship-supported and policy-only routes, about real episodes instantiating the joint architecture, and about whether grouping judgments by their contributing organization yields explanatory gains. Retaining those hypotheses does not imply that their antecedents can currently be measured adequately.

Before a proposed route experiment could become a test, researchers would need an independently justified way to identify the candidate routes, determine what an intervention changes, and distinguish continuing contribution from a dormant alternative. Redundant support makes the problem harder: two routes may each help form a judgment even when removing either one would leave the same spoken conclusion. Changing confidence or speed is not sufficient evidence of the relevant contribution unless the connection has been defended independently.

Systems whose internal computations can be inspected may permit more explicit tests of a supplied architecture. Even then, observing a computational dependency does not by itself settle what a representation means or whether the structure is an adequate model of human moral cognition. The improvement is access to the implementation, not exemption from interpretation.

Where methods are inadequate, the appropriate result is a clear statement of the identification problem. Where independently validated methods eventually support a discriminating comparison, adverse results must be allowed to count against the applicable hypothesis. Present inability to perform that comparison should not be advertised as its successful completion, nor as proof that no future evidence could matter.

5. What would change the assessment?

A conceptual objection could show that a central distinction collapses under careful examination, that the proposed moral boundary depends on moral classification in disguise, or that a rival preserves the same distinctions with less or better structure. A formal objection could expose a counterexample to a claimed consequence. Neither objection has to wait for a brain-imaging method.

A stronger empirical result could show that an association survives matched measurement, new samples, competing explanations, and tests on new cases. A causal study could establish the effect of a particular manipulation without establishing every proposed intermediate process. A genuinely validated identification method could make more demanding claims about contributing organization assessable. Those would be different advances, and the conclusion should name the one actually achieved.

Negative evidence also needs a target. Failure of a questionnaire to measure agency does not show that agency is absent. Failure of one numerical interaction does not erase the distinction between positive and negative inherent value. Conversely, repeated failure under adequate, independently specified methods cannot be indefinitely explained away by inventing a new hidden route for every inconvenient case. The formal account therefore requires prospective comparison rules and permits unresolved outcomes; it does not license fitting the definition to the desired verdict.

The represented-subject condition is a particularly clear example. NDT says that an event can be terrible without being immoral when no agent is represented as the subject of the appraisal. A critic may challenge that proposed boundary conceptually. An empirical argument for subject absence would additionally need evidence capable of distinguishing absence from failed detection, including implicit representation. No such detector is supplied by the theory. Merely asking whether the event is “morally bad” does not resolve either dispute.

Conclusion

NDT asks readers to examine the structure of judgment: what someone understands, what matters within that understanding, and how a conclusion is formed. The power of that proposal, if it succeeds, lies first in making distinctions intelligible and showing how they fit together. Formalization exposes commitments that can then be tested for coherence and compared with alternatives.

The completed studies provide a narrower empirical contribution. They show reported associations between issue-specific beliefs and controversial judgments that were stronger than those of the measured broad values in the authors' analyses. They motivate closer attention to the understood situation. They do not reconstruct cognitive provenance or confirm the full relationship architecture.

The next steps should preserve both kinds of work. Develop the argument and its comparisons; improve the empirical tests that are presently feasible; state the missing methods wherever a stronger claim depends on them. A theory can be worth understanding before every application can be measured precisely. That is a reason to explain its argument well, not a reason to claim evidence that does not yet exist.

Sources and technical continuation

Beal, B., & Rottman, J. (2024). *Nonmoral beliefs are better predictors of moral judgments than moral values*. Submitted manuscript, project-held copy. [Read the manuscript](#).

For exact definitions, consult [Formal Specification 1.6.4](#), especially the moral-judgment criterion, worked composition tests, contribution conditions, and evidence sections. [Scope and Commitments 1.17](#) explains the intended boundaries. [Proposition Registry 2.10](#) distinguishes assumptions, constitutive proposals, formal consequences, and empirical hypotheses. [Empirical Audit 2.4](#) contains the fuller catalogue of proposed investigations. Descriptions of a developed identification program in that catalogue name the specification of a research question, not a validated instrument or demonstrated capacity to read exact human reasoning routes.