

From Representation to Obligation

A General Architecture of Normative and Moral Cognition

Bree Beal Reader-facing article, Draft 0.23 **Status:** Unratified RC3 author draft; not an approved theory artifact.

Abstract

Normative Detachment Theory (NDT) explains how someone forms prescriptions and appraisals from a representation of a practical situation. It takes the relevant representations, resources, and candidate routes as supplied inputs; it does not yet explain how perception or memory generates them online. NDT begins from a simple point: cognition need not begin from an evaluatively blank representation of the world. A representation may already distinguish success from error, danger from safety, or an outcome to be preserved from one to be avoided. A working model of the situation then connects this evaluative structure to possible actions, expected outcomes, requirements, priorities, and challenges. When an adequately supported conclusion forms, an “ought” has detached from an evaluatively structured representation of what is.

Moral cognition is a narrower case within this general architecture. A normative judgment is moral when a represented moral relationship helps organize the evaluation that actually produces that occurrence of judgment. NDT now describes that relationship through three independently variable parts: an inherent-valence profile, Relational Positioning, and perceived agency. Positive and negative inherent value are separate components, so a mixed profile is not reduced to a net score. Relational Positioning locates this exact object relative to this exact subject through identity-mediated positioning, intimacy, or direct encounter. Identity-mediated positioning can be shared, contrastive, or complementary-role based; an institutional role such as physician relative to patient is one possible complementary form when it is operative in the reasoning. Mere category membership, role possession, or physical proximity is insufficient. Functional access and actual contribution remain separate from the relationship core.

A representation can also affect what the relationship requires, permits, prioritizes, or appraises without helping constitute the core. NDT calls that downstream, core-preserving operation a strict bender. The same representation may support separate valence, positioning, bender, or frame-attractor edges, but each edge must be credited and tested independently.

Because agents, objects, actions, and relationships are represented under interpretations, the same case can sustain several locally coherent judgments. When those judgments become comparison-ready and conflict, NDT calls the resulting many-voiced structure *normative polyphony*. Together, evaluatively structured representation, normative detachment, relational moral provenance, and frame-relative polyphony form a unified account of how cognition moves from a meaningful world to normative and moral judgment.

1. Seven explanations of one act

Near the end of *Crime and Punishment*, Raskolnikov confesses his murder to Sonya. Yet the confession does not arrive as one stable account. He redescribes what he did repeatedly. In one telling, he acted to protect his family. In another, he was a petty thief. Elsewhere, the murder appears as a response to desperate social conditions, an act of spite against a corrupt world, a test of whether he belongs among the extraordinary, or an attempt to escape inherited morality. At one point he says, in effect, that the person he killed was himself (Beal, 2018).

The event does not change. The practical and moral world in which it appears does. Different descriptions foreground different agents, objects, relationships, motives, possibilities, and questions. Under one interpretation, the pawnbroker appears as an exploiter and the act as rescue. Under another, she appears as a person murdered for money. Under another, the relevant object is Raskolnikov’s own identity, and the crime becomes self-destruction. Some of these descriptions support conclusions that genuinely contradict one another. Others answer different questions and therefore pass without meeting.

Dostoevsky magnifies a structure found in ordinary judgment. The same intervention can be represented as medical care, bodily violation, professional duty, paternalism, mercy, or abandonment. The same person can appear as a stranger, fellow citizen, rival, family member, threat, patient, or bearer of rights. Two people may invoke justice while disagreeing about who counts as the victim; invoke care while construing the relevant relationship differently; or invoke dignity while disagreeing about what kind of being or action is before them. Even one person can move among several coherent interpretations of the same episode.

This creates a difficulty for theories that begin with visible moral products. Moral words do not identify the cognition that generated them. Prosocial behavior can arise from concern, habit, self-interest, reputation management, fear, or institutional discipline. Rule-following can be moral, legal, prudential, ritual, or merely automatic. A utility calculation can express

concern for affected people, but it can also be a resource-allocation routine that represents no one as morally significant. An action can match a moral rule while the thought producing it is not moral in kind.

The reverse problem also occurs. A morally organized judgment may be expressed in ordinary language, formed rapidly without conscious reasoning, or defeated by weakness of will before it becomes action. If moral cognition is identified with a special vocabulary, philosophical doctrine, emotion, topic, or behavior, both kinds of case become difficult to describe.

Normative Detachment Theory begins earlier. It asks how someone represents a practical situation, how that representation becomes action-guiding or appraisal-guiding, and what additional structure makes one normative judgment moral while another remains prudential, institutional, aesthetic, epistemic, or instrumental. The unit of explanation is not a sentence or behavior considered in isolation. It is one occurrence of judgment, formed under an interpretation and through a particular route of support.

2. What is normative detachment?

Normative detachment is the formation of a prescription or appraisal from a represented practical situation. The term comes from logic: when relevant conditions are satisfied, a conclusion becomes available, or “detaches.” In NDT, the conclusion is a judgment such as *I should do this, that action violated a requirement, what happened was terrible, or the agent acted wrongly*. The current architecture begins after its representations and inferential resources have been supplied; an end-to-end account of their online generation is outside its present scope.

Consider a nonmoral example. You are assembling a chair. You have taken completion of the chair as a goal. You represent tightening a screw as necessary to complete it and turning the screw clockwise as the available means. If no stronger requirement blocks the act, the practical conclusion is that you should turn the screw clockwise, relative to that goal. If the screw is reverse-threaded, the represented action–outcome relation changes, and so does the conclusion.

The example is simple, but it displays the basic form. The person represents:

- a practical situation and a point of view within it;
- possible actions;
- expected consequences of those actions;
- an evaluative ordering, goal, or standard;
- requirements and exceptions;
- priorities among considerations; and
- support for a particular prescription or appraisal.

NDT calls this organized working picture a *decision model*. The phrase should not suggest a conscious spreadsheet in the mind. A decision model is a functional organization: the practical problem as the person is currently representing it. Its implementation may be deliberate or intuitive, symbolic or neural, linguistic or perceptual, or some mixture of these.

Detachment is about judgment formation, not behavior. A person may judge that an action is required and fail to perform it because of fear, competing motives, weakness of will, incapacity, coercion, or circumstance. Conversely, a person or machine may perform the prescribed action through habit, punishment avoidance, or script completion without forming the relevant judgment. The architecture therefore stops at the judgment event. Action lies downstream.

Normative judgment is also broader than moral judgment. A proof can be elegant, a golf swing incorrect, a machine miscalibrated, a choice imprudent, an institutional rule binding, or an event tragic without the judgment being moral. Moral cognition is not a separate faculty added alongside these forms. It is a more specifically organized subset of normative cognition.

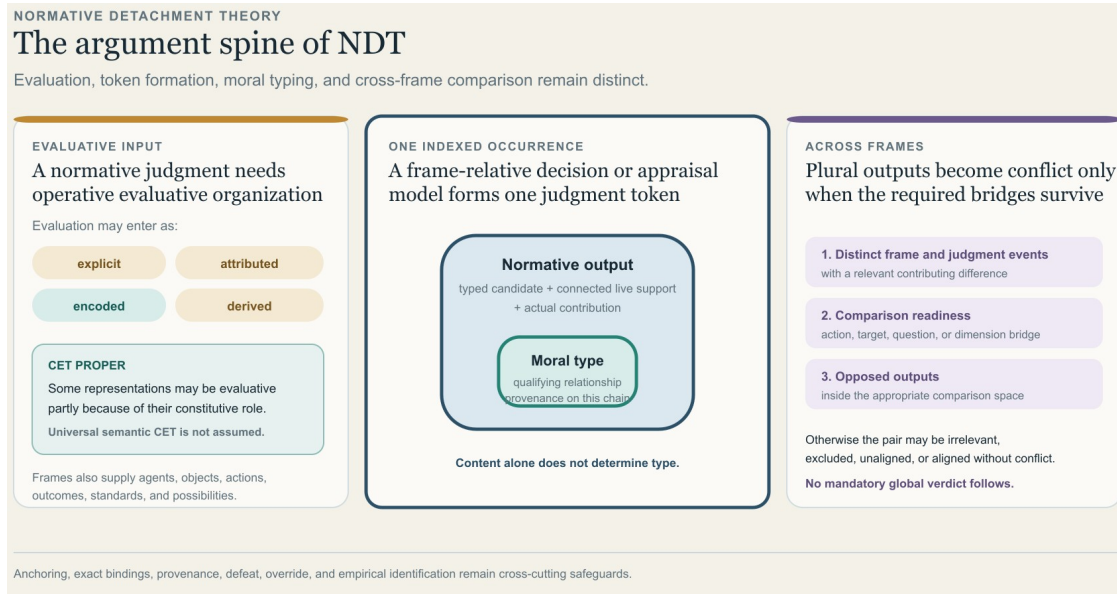
NDT develops this proposal through three connected theses:

1. **The Constitutive Evaluativity Thesis (CET):** some active representations are evaluative partly because of what they represent or because of the functional role that makes them the representations they are.
2. **The Normative Detachment Thesis (NDT):** a framed action-guiding model connects evaluative organization to possible actions, requirements, appraisals, priorities, and surviving support, allowing a normative judgment to form.
3. **The Moral Detachment Thesis (MDT):** a normative judgment is moral when a qualifying represented relationship helps organize the evaluation that actually produces that judgment.

The first thesis explains why cognition does not always begin with a value-free description to which a separate desire must later be attached. It does not claim that encoded evaluation is the only source of evaluation. The broader evaluative prerequisite is that every normative detachment needs some operative evaluative organization, which may be explicit,

attributed, encoded, or built from those sources. The second thesis explains why evaluative information alone does not yet determine a specific ought. The third explains what distinguishes moral from nonmoral normative cognition.

The Explorer organizes this argument into four explanatory axioms: **Actual Framing**, **Evaluative Grounding**, **Normative Detachment**, and **Relational Moral Type**. That four-part organization is a reader-facing compression of several governing clauses, not a claim that four sentences logically entail every definition, parameter, and implementation choice in the full specification. The four axioms organize the constitutive spine; definitions, scope conditions, bridge principles, empirical parameters, and replaceable implementations supply the rest of the formal apparatus.



3. From an evaluatively structured “is” to an “ought”

NDT describes how ought is inferred from is in cognition. It does not claim that a prescription follows from a premise stipulated to be empty of every value, goal, standard, priority, or evaluative ordering. The relevant “is” is the situation as someone actually represents it while forming a judgment, and cognitively active representations need not be evaluatively empty.

Three kinds of ordering

The theory distinguishes three structures that can be run together.

First, a representation has conditions of success and failure. A perceptual state may track something accurately or inaccurately; a categorization may fit or misclassify; a signal may be relevant or noise. This is a thin semantic ordering. It distinguishes better and worse representational performance, but it does not yet rank possible states of the world or tell an agent what to do.

Second, someone may treat represented outcomes differently. Some outcomes are maintained, approached, restored, or stabilized; others are avoided or corrected. A threat representation, for example, is not merely a neutral picture followed by an arbitrary negative label if sensitivity to danger helps constitute its cognitive role. This is outcome evaluation.

Third, a decision model connects outcome evaluation to feasible actions and applicable requirements. Accuracy alone never tells an agent whether to preserve a specimen, flee a fire, record a measurement, or remain still. An action-guiding ordering depends on what outcomes are represented, what actions are possible, what standards apply, and what constraints survive.

The transitions are therefore:

representational success and failure → represented outcome evaluation → frame-relative guidance or appraisal

Each arrow does genuine work. Semantic correctness does not directly select an action. Outcome evaluation does not by itself determine which action is feasible or required. An action-guiding model must connect evaluation to a particular agent, situation, set of possibilities, and body of standards.

Where evaluation comes from

Evaluation can enter cognition in several ways.

It may be **explicit**, as when a goal, desire, command, policy, or principle is represented separately. This includes familiar belief–desire reasoning: I want the chair completed; I believe turning the screw is necessary; therefore I judge that I should turn it.

It may be **attributed** by a separable process. A sensor can supply a neutral representation to a distinct controller that assigns importance or reward.

Or it may be **encoded** in the representation's role. A state may count as representing danger partly because it feeds a standing organization disposed to avoid, correct, or stabilize the relevant outcomes. In this case there need be no separately represented sentence such as *I desire to avoid danger*. The difference between outcomes is built into the way the representation functions.

Encoded evaluation is CET's distinctive case. It is not the claim that semantic accuracy contains a moral rule, that every representation is an imperative, or that the theorist may impose a preferred welfare ranking. It is the claim that some representations are partly constituted by their place in a control-sensitive organization. Whether this is common, how it should be measured, and where encoded evaluation ends and a separable evaluator begins are empirical questions. The conceptual point is that a cognitively instantiated description may carry evaluative structure without containing an explicit desire-token.

The relation to Hume

Hume's familiar is–ought passage has a familiar reading on which it prohibits a logical derivation of prescriptions from purely descriptive premises (Hume, 1739–1740/2000). NDT accepts that boundary. If the starting model is stipulated to contain no evaluative ordering, goal, requirement, priority, or commitment, NDT produces no ought. Nothing normative appears from nothing.

The theory questions a different assumption: that the representations from which action-guiding cognition begins must be empty in this way. A representation of spoiled milk can distinguish a condition the person is organized to avoid. A representation of a threat, opportunity, error, injury, or promise can already occupy a role structured by what someone preserves, corrects, approaches, or treats as significant. NDT traces the additional steps by which this evaluatively structured representation participates in a prescription or appraisal.

The claim is therefore straightforward and limited: **NDT describes how someone can infer ought from an evaluatively structured is**. It neither derives value from value-free propositions nor reduces normativity to semantic correctness.

From a candidate judgment to a formed judgment

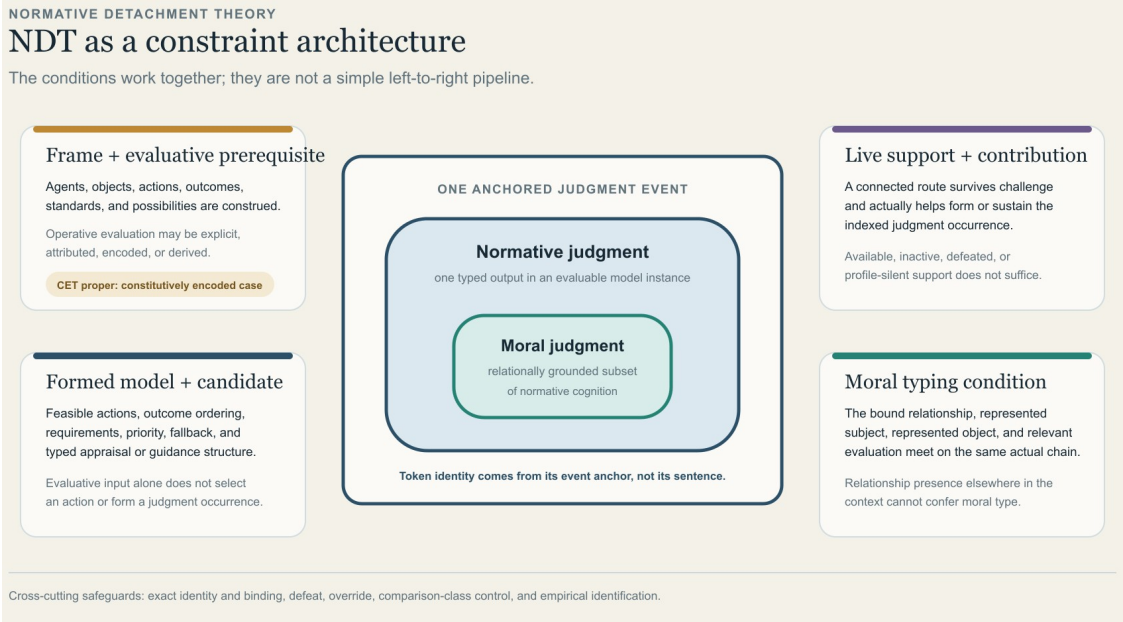
A decision model can make a conclusion available before it becomes the person's judgment. NDT calls the available conclusion a *candidate*. A candidate becomes an output only when a connected route of support remains active and actually helps form it.

This distinction matters because cognition contains more than the reasons it uses. A moral relationship, policy, fear, professional rule, memory, and anticipated consequence may all be active at once. Their availability does not show which of them produced this occurrence of judgment. NDT therefore distinguishes the background field from *actual contribution*.

Actual contribution is more demanding than a line one can draw after the fact. A source must participate in an active sufficient process and make a difference to some relevant feature of the judgment-forming event. For the relationship profile specifically, the difference must be either a controlled evaluative effect on the same route that forms the judgment or a minimal supporting role within a sufficient profile on that route. Merely changing the profile alongside an unrelated feature is not enough, and an effect on one route cannot be used to certify contribution on another. Several routes can still contribute redundantly even when removing either one would leave the same words intact. Conversely, when evidence cannot distinguish a contributing route from an available backup, the token's provenance may remain unresolved.

Support can also be defeated. A represented objection may challenge a premise, an inferential link, or the formation of a conclusion. But a consideration does not block a route merely because an outside observer can imagine it. The person forming the judgment must register the consideration, connect it to a target, and treat it as relevant. NDT's formal implementation uses a conservative defeat rule, while allowing that actual cognition may instead employ graded, probabilistic, or more trusting forms of challenge handling.

Requirements and priorities remain visible through this process. If a stronger consideration licenses an exception, the overridden requirement can leave residue: regret, apology, repair, or recognition that something of value was still violated. If every feasible option violates an important requirement, a person may be guided toward the least serious violation without pretending that the selected act became wholly acceptable. This is how NDT represents tragic choice without collapsing it into simple outcome maximization.



4. Moral type is a matter of actual provenance

The movement from representation to normative judgment does not yet explain morality. A safety system can avoid spoiled food. A chess engine can distinguish stronger from weaker moves. An institution can apply a rule. Each performs evaluatively organized cognition in a broad sense. What makes one normative judgment moral?

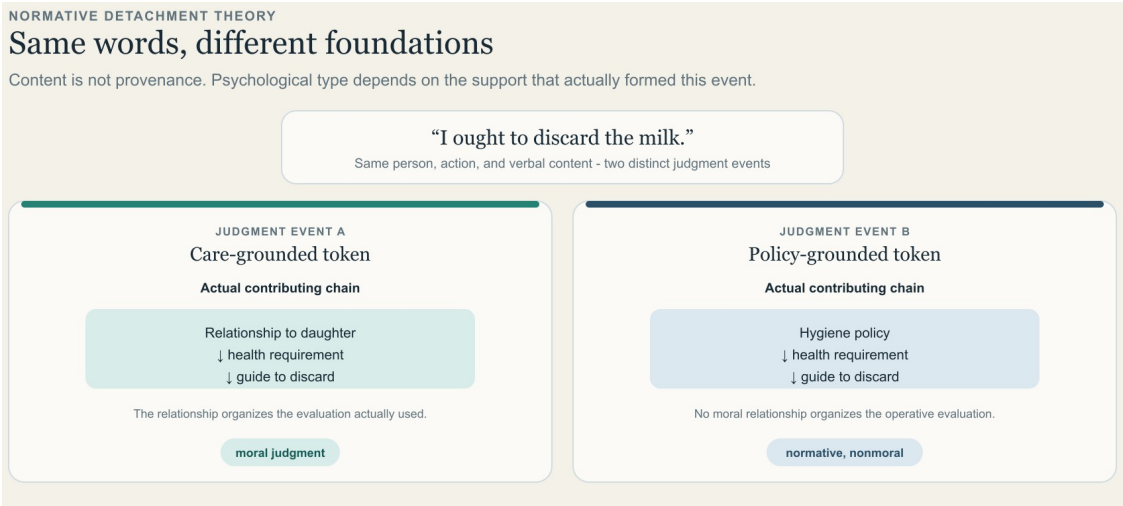
The Moral Detachment Thesis gives a source-sensitive answer:

A normative judgment is moral when a represented moral relationship, with the relevant subject and object, helps organize the evaluation that actually forms that judgment.

The relationship must not merely occur somewhere in the situation. It must contribute on the same connected route that supports the operative requirement, ordering, acceptability boundary, priority, or appraisal.

Consider a parent deciding whether to use milk that smells spoiled. In one occurrence, the parent represents the child’s health as mattering within their relationship. That relationship helps establish illness as unacceptable, and the judgment forms: *I should discard the milk*. In another occurrence, the same person reaches the same conclusion solely by applying a kitchen-safety policy learned at work. The child and the parent’s care may remain in the background, but they do no work in this judgment.

The sentence, agent, available actions, and outward behavior can be identical. The first judgment is moral if the relationship actually shapes the evaluation. The second is normative but need not be moral. This is not a difference between good and bad outcomes, sincere and insincere people, or conscious and unconscious reasoning. It is a difference in the functional provenance of two judgment events.



The same principle applies to explicit philosophical reasoning. A universalization test, utility calculation, religious command, or rights principle can function inside moral cognition when a represented moral relationship helps organize its evaluation. The same procedure can also operate as institutional policy, strategic calculation, intellectual exercise, or habit. Deontological or consequentialist vocabulary does not determine the type of every thought occurring under that description.

NDT therefore classifies tokens rather than strings. A *token* is one judgment occurrence anchored to one judgment event. Its content, active interpretation, action-guiding model, and actual support characterize how it was formed; they do not jointly individuate the token. Two tokens can contain the same words while belonging to different cognitive kinds. One token can also contain several converging sources, moral and nonmoral, provided each genuinely contributes.

This source-sensitive criterion prevents two false positives. First, conformity to a moral rule does not prove that moral cognition produced it. Second, the mere presence of a morally significant person or relationship does not moralize every judgment formed in the same context. A care relationship can motivate attention while a policy alone produces the conclusion; a moral object can be represented on one inactive route while fear produces the operative judgment. Moral type requires relationship-organized evaluation on the actual chain.

The theory is descriptive. To classify a judgment as moral is not to endorse its content, excuse the agent, or assign moral worth to the action. A judgment can be moral in cognitive type while another person, community, or ethical theory rejects it as cruel or mistaken. Conversely, an outwardly admirable action may have been generated through nonmoral cognition.

5. The relationship at the center of moral cognition

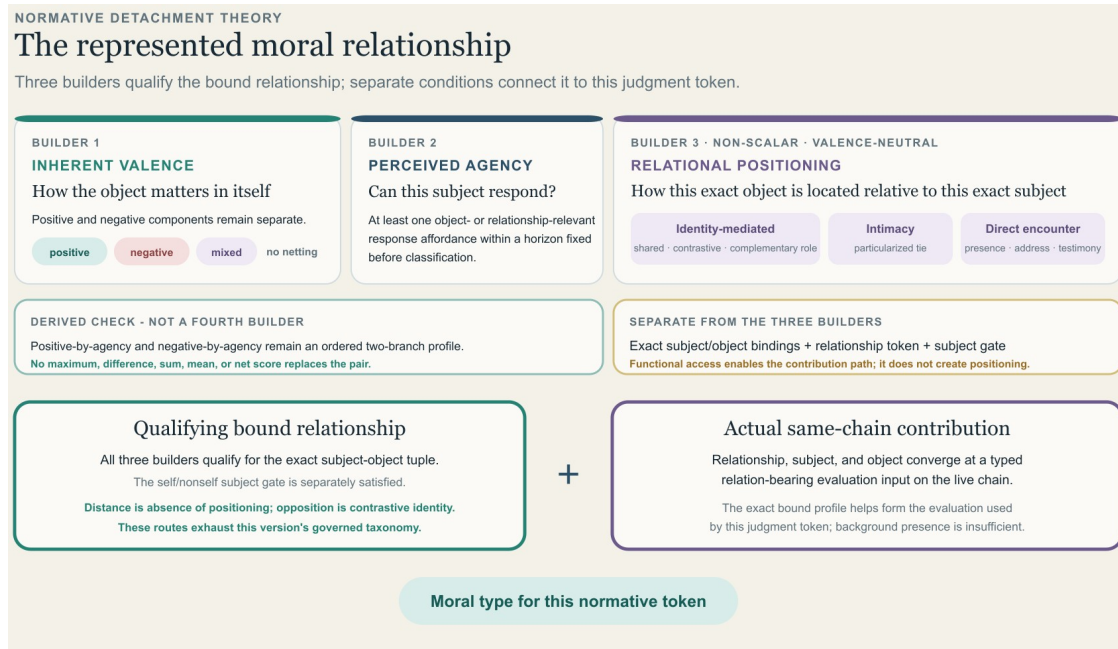
The phrase *moral relationship* would be circular if it meant only “whatever relationship produced a moral judgment.” NDT instead proposes a typed structure that can be specified before the target judgment is classified.

The current relationship core has three independently variable parts:

- **Inherent valence:** the object is represented as mattering in itself, positively, negatively, or both, rather than only as a means. Positive and negative components remain separate. A mixed profile therefore records ambivalence or conflict instead of cancelling to zero; a profile with neither component lacks this builder.
- **Relational Positioning:** a representation locates this exact object relative to this exact subject in the active situation model. It is not a general similarity or distance score.
- **Perceived agency:** within a declared present, prospective, retrospective, or counterfactual horizon, the subject is represented as able to affect, respond to, answer to, or otherwise treat the object.

NDT does not set one universal numerical threshold for these roles. Its finite executable companion uses an explicitly replaceable gate so that positive-only, negative-only, mixed, and neither profiles can be tested without allowing positive and negative components to cancel one another. Calibrating any population or system requires independently justified indicators and decision rules.

The companion also keeps separate positive-valence-by-agency and negative-valence-by-agency branches before a profile-preserving derived relationship-strength gate, which is not a fourth builder.



Three ways of positioning an object

Relational Positioning has three governed routes. **Identity-mediated positioning** uses an operative identity or role schema as a bridge between subject and object. Its forms are shared identity (*one of us*), contrastive identity (*one of them* or *my counterpart*), and complementary role (*physician–patient*, *guardian–ward*, or another paired role). A label or category is not enough: the schema must locate the exact represented object relative to the exact represented subject and contribute through that bridge.

Intimacy positions the object through a particular relationship—shared history, trust, dependence, mutual recognition, disclosure, or relationship-specific knowledge. **Direct encounter** positions the particular other through presence, expression, testimony, address, condition, or vulnerability. These routes can overlap. None is individually necessary, and mere physical distance is not a fourth route: under an adequate scheme, distance is the absence of an operative route. Likewise, opposition is not negative inherent value. Opposition belongs to contrastive identity-mediated positioning; negative valence states how the object is evaluated.

The former “standing-mediated positioning” category is retired. Cases it was meant to cover divide more precisely. An institutionally directed assignment can be complementary-role identity-mediated positioning. A general status or rule that changes what follows from an already represented relationship belongs downstream as a bender when it passes the bender test. Functional access remains enabling machinery: an otherwise complete relationship cannot affect this token through a blocked process, but access alone does not create Relational Positioning or moral type.

Core and downstream effects

NDT reserves **strict bender** for a downstream operation that preserves the exact relationship-core snapshot while changing one fixed downstream locus through the required live contributing path. If changing the representation instead changes inherent valence, Relational Positioning, or perceived agency, it is a builder input, realizer, or modulator—not a strict bender.

This boundary prevents double counting. A physician identity may position physician and patient through a complementary role. A separate edge may use that identity to support confidentiality, urgency, or a duty of care. The first edge helps constitute the relationship; the second bends what follows from it. One representation may support both, but the evidence for one edge cannot be recycled as evidence for the other. The same independence rule applies to an attractor that merely makes a frame more likely to be selected.

What removal reveals

Removal tests target different claims:

- Remove both inherent-valence components while holding positioning and agency fixed: the model loses the distinction between an object represented as inherently significant and one represented only instrumentally.

Removing only one component tests positive-only, negative-only, and mixed profiles without netting them together.

- Remove Relational Positioning while holding valence and agency fixed: the model loses the distinction between abstract evaluation of an object and locating this object relative to this subject.
- Remove one positioning route or identity form: only that bridge and its transfer predictions disappear; another route may still position the object.
- Remove perceived agency: the model loses the represented capacity of this subject to affect, answer to, or respond to this object.
- Block functional access: the relationship core may remain represented, but that process cannot contribute to this judgment.
- Remove a strict bender while holding the core fixed: the relationship remains, while the downstream requirement, permission, priority, appraisal, or responsibility effect should change.

These tests defend conditional distinctions inside NDT's current decomposition. They do not prove that its labels are uniquely correct, that its finite gates are universal, or that every rival must use the same modules. A rival can redistribute the work. The challenge is whether it can preserve the distinctions without quietly rebuilding equivalent structure.

The represented-subject gate remains separate. Perceived agency helps constitute the relationship; the gate asks whether a nonself subject is represented as able to recognize and integrate the relevant standing in a response. Responsibility and blame require further conditions. Entity kind does no formal work by itself.

6. Moral subjects, moral objects, and the grammar of appraisal

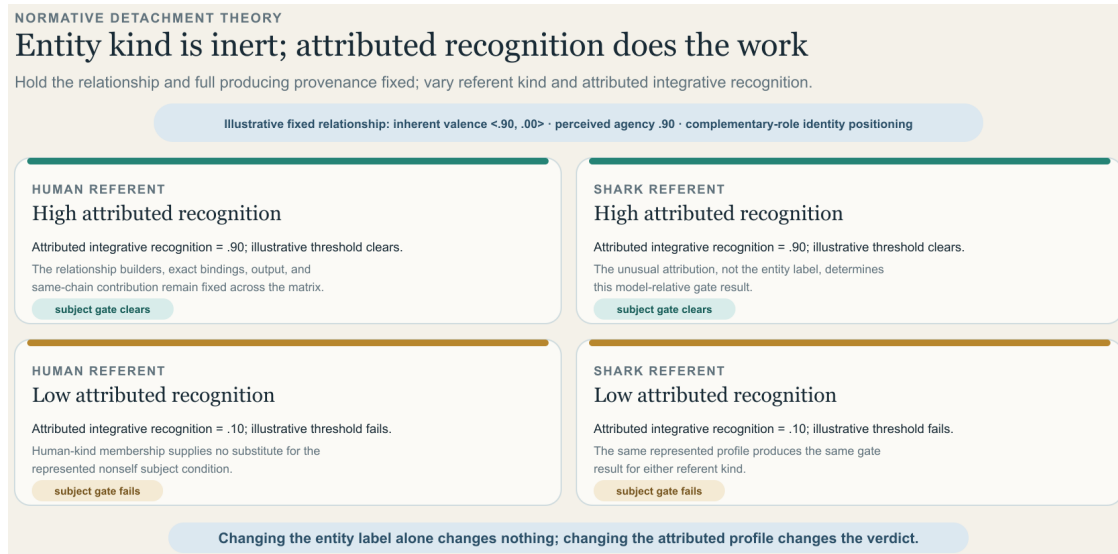
A moral judgment contains at least one represented moral subject and one represented moral object, though either may be implicit in the sentence. The subject is the person, collective, institution, self, or other agent represented as acting, responding, refraining, preserving, cultivating, repairing, or standing under a norm. The object is what is represented as bearing the relevant standing, welfare, condition, treatment, or interest.

The roles can refer to different entities or to the same one. In self-directed judgment, the judging person can be both subject and object without the roles collapsing. *I am the one acting* and *my own standing is at issue* are different facts. Self-reference clears the special third-person subject gate, but it does not automatically establish inherent valence, Relational Positioning, perceived agency, functional access, a qualifying bound relationship, or profile-sensitive actual contribution. NDT can therefore distinguish moral concern for oneself from prudential self-management, self-alienation, or a policy applied to oneself without moral provenance.

Surface grammar can conceal the subject. *What happened to her was wrong* may be a moral judgment when the actual derivation represents a responsible official, institution, or person linked to the event. A judgment that an institution is corrupt can be moral when the institution, or members acting through it, is represented as a collective or distributed agent.

A genuinely agentless appraisal is different. An earthquake represented as having no agent behind it can be terrible, tragic, catastrophic, or sacred. A relationship to its victims can ground obligations to rescue, mourn, rebuild, or prepare. Yet the earthquake itself has not acted immorally unless some person, collective, institution, or other agent is represented as responsible for the event or its consequences. NDT's broad appraisal grammar preserves negative evaluation without forcing every terrible occurrence into wrongness or blame.

This boundary is an explication claim, so a lexical verdict alone cannot settle it. But a robust judgment episode for which a validated detector supplies strong negative evidence against explicit and implicit subject representation counts as evidence 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. The inference must report the detector's sensitivity, false-negative rate, temporal window, and decision threshold rather than treating a failed detection as proof of absence. A rival that preserves the surrounding distinctions strengthens the challenge but is not required before the evidence becomes adverse.



Moral objects are equally unrestricted by entity kind. A person, animal, ecosystem, institution, artifact, abstract class, future generation, ideal, or way of life can occupy the object role if the relevant builder profile and relationship are represented. No label—*humanity, nature, rational agency, sentient welfare, the nation, the sacred*—automatically supplies the relation. The person forming the judgment must actually construct it within the frame and use it in the judgment.

NDT also distinguishes the relationship core from downstream effects. A role such as physician, guardian, citizen, or officeholder may supply complementary-role identity-mediated positioning when it locates the exact object relative to the exact subject. Any requirement, permission, priority, appraisal, or responsibility consequence supported by that identity must be represented separately as a strict-bender edge. Power, vulnerability, sentience, authority, and similar features are candidate sources, not automatic benders. They count as strict benders only when the core remains fixed and a registered downstream locus changes. Attractors remain different again: they alter which frame is likely to become active without serving as reasons within the selected frame.

7. Framing comes before the application of values

A value does not determine its own objects. Justice cannot say, by itself, who counts as the claimant, victim, offender, or rightful owner. Care does not identify the being toward whom care is owed. Dignity does not determine which entity is represented as a person or what treatment counts as degrading. Before a principle can be applied, cognition has already organized a world of agents, objects, actions, identities, episodes, relationships, and possibilities (Beal, 2020).

NDT calls that coordinated interpretation a *frame*. A frame is not merely a verbal label. It is an active construal that fixes what the case is, who is involved, what actions are available, which outcomes are relevant, and what question is being asked. The same thin referent—roughly, “that person” or “that occurrence”—can be construed differently across frames.

This explains why shared moral vocabulary can fail to locate disagreement. Two people may both value care yet disagree because one represents a fetus as a child and another as developing tissue. Both may value loyalty while construing the relevant group differently. Both may condemn violence while disagreeing about whether one represented execution is attack, defense, punishment, treatment, or withdrawal of assistance. The disagreement can occur beneath the level measured by a list of abstract values.

Framing also clarifies moral exclusion. A person can disappear from moral concern not only because an explicit value is rejected, but because the person is represented as a threat, statistic, tool, contaminant, outsider, or nonagentive object. Such a change can alter inherent valence, remove or reorganize Relational Positioning, block functional access, weaken the action-relevant connection, or alter attributed moral vision before an explicit principle is ever weighed. Dehumanization is therefore not merely a belief that some rule permits mistreatment. It can involve a reorganization of the perceived moral relationship itself.

Frames are neither automatically accurate nor equally defensible. They can be distorted, strategically induced, evidence-resistant, or defeated. NDT’s descriptive neutrality does not imply relativism about truth or moral correctness. It says that a theory of cognition must represent the interpretation that actually organized the judgment, including interpretations the theorist regards as false or abhorrent.

Because frames determine action descriptions, two judgments cannot be declared contradictory merely because their sentences look opposed. A physician’s act may be described as *ending treatment* in one frame and *honoring a refusal* in another. These descriptions may map to the same physical execution, but the appraisals may still answer different questions.

NDT therefore requires represented bridges across frames: a link between the relevant actions or targets and a common evaluative question or comparable dimensions.

This discipline prevents an observer from manufacturing conflict by treating unlike appraisals as if they occupied one scale. It also explains how conflict can emerge when a new bridge becomes available. Two judgments that previously concerned different acts or questions may become comparison-ready when the person recognizes them as accounts of the same execution and as answers to a shared normative question.

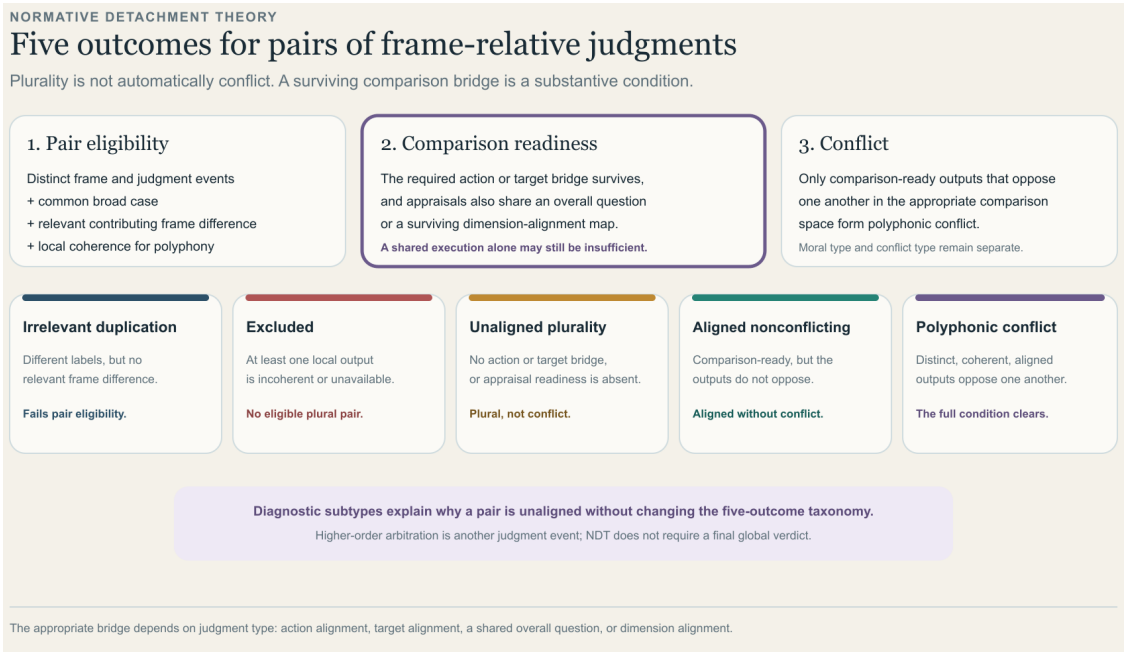
8. Normative polyphony

Several interpretations of one case can remain active without being merged into a single verdict. When distinct frames support locally coherent judgments, when the relevant frame differences matter to those judgments, and when comparison-ready outputs conflict, NDT calls the structure *normative polyphony*.

The term comes from Mikhail Bakhtin’s account of Dostoevsky (Bakhtin, 1929/1984). In a polyphonic novel, voices are not merely fragments waiting for the author to announce the one correct synthesis. Each retains an organized perspective. NDT extends this idea to cognition. A person can sustain several framed normative routes whose conclusions remain intelligible from within their own action-guiding understandings (Beal, 2022).

Polyphony is not a synonym for inconsistency. It differs from:

- a tie between equally ranked acts inside one decision model;
- tragic choice in which every option violates a requirement;
- simple indecision;
- duplicate expressions of the same judgment;
- several judgments that concern different cases; and
- plurality that has not been made comparison-ready.



Raskolnikov’s confession is a limit case because the frames shift repeatedly while remaining anchored to one deed. Defender, thief, victim, rebel, visionary, and self-destroyer are not interchangeable labels pasted onto one fixed description. Each reorganizes motives, relationships, possible alternatives, the object of appraisal, and the question the act is supposed to answer. Some pairs can be aligned and conflict. Others remain unaligned because one concerns social rescue, another personal greatness, and another the destruction of the self.

Polyphony need not be a defect. Sometimes it reveals that a situation contains several real relationships or legitimate questions that cannot be reduced without loss. At other times, one frame is poorly supported or strategically evasive. NDT does not infer equal validity from local coherence. It identifies the many-voiced organization so that the sources of conflict, evasion, or insight can be examined rather than hidden behind one verbal verdict.

Nor does the theory require a final global judgment. A higher-order frame may compare lower-order outputs and form an all-things-considered conclusion. But that arbitration is another judgment event with its own interpretation and support. It

can itself be challenged or stand in conflict with lower-order judgments. Cognition may reach closure; NDT does not build closure in as a necessary endpoint.

This account helps explain why reflection can increase uncertainty rather than eliminate it. Deliberation may reveal new consequences or settle a factual question, but it can also make another relationship or self-construal visible. Dialogue can produce convergence by aligning frames, or deepen conflict by showing that the parties inhabit different representations of the case. The mechanism lies not merely in possessing several values, but in the availability of several action-guiding understandings to which values and requirements can be applied.

9. What NDT changes in the study of morality

NDT reorganizes familiar phenomena rather than denying their importance. Harm, fairness, cooperation, norms, emotion, character, mind perception, care, welfare, rights, and sacredness all matter. The question is where they enter the architecture and what they can establish.

Some theories identify recurring contents of moral concern. These contents can function in NDT as appraisal dimensions, requirements, outcome structures, relation qualifiers, or culturally common frames. They do not by themselves define moral type, because the same content can be represented in a legal, strategic, prudential, institutional, or moral route (Beal, 2020, 2021).

Some theories model reasoning procedures: utility maximization, universalization, bargaining, rule application, practical syllogisms, reinforcement learning, Bayesian inference, or defeasible argument. NDT can represent a procedure inside a decision model without inferring moral type from the procedure alone. A procedure becomes part of moral cognition when a qualifying relationship contributes to the evaluation used by the token.

Some approaches focus on outward signatures: authority independence, generalizability, seriousness, punishment, emotional force, or characteristic language (Turiel, 1983). Such measures can provide evidence about moralization. They remain defeasible because a trained person or artificial system can reproduce the signature through instruction, imitation, policy, or association. The same limitation applies to cooperative behavior. Cooperation is one important context for morality, but reputation management, habit, fear, and self-interest can produce it as well.

Other theories contribute formal tools. Work on default reasoning and formal argumentation helps NDT represent exception and challenge (Pollock, 1987; Dung, 1995; Horty, 2012); work on causal contribution and deontic detachment helps discipline redundant support and the movement from conditional requirements to actual judgments. NDT's further question is what constructed the reasons and relationships entering those tools, and which sources actually contributed to this token.

The distinction between behavior and provenance has ancient and diverse antecedents. Kant's prudent shopkeeper is one familiar illustration: outwardly honest conduct may lack the motive relevant to moral worth (Kant, 1785/2012). Contemporary work on moral worth, right reasons, and virtue develops related questions about credit, motive, responsiveness, and character (Herman, 1993; Arpaly, 2003; Markovits, 2010; Sliwa, 2016). NDT's target is narrower and different. It classifies the cognitive type of a judgment event rather than deciding how much moral credit an action or agent deserves. Moral worth can vary while NDT type is held fixed, and NDT type can vary while behavior is held fixed. The shopkeeper illustrates a broad source-sensitivity insight; it neither supplies NDT's token-provenance architecture nor establishes a special Kantian lineage for an idea expressed across moral, religious, and philosophical traditions.

Character appraisal also belongs within the architecture without replacing it. Third-person and first-person retrospective judgments may use behavior as evidence about character: *Was I acting from care or merely following policy?* Such cases can be modeled as agent-directed appraisal and may require probabilistic relations among character, expected conduct, observed behavior, and judgment provenance. First-person prospective cases remain especially useful for distinguishing the models. When a person is deciding what to do before acting, the proposal that every relevant judgment is merely an inference about the person's character becomes strained. NDT treats character as one possible target and source of evidence, not as the universal object of moral cognition.

Artificial systems sharpen the same questions. A system can reproduce moral labels, follow safety rules, calculate welfare, and generate persuasive explanations. None of these abilities alone establishes moral cognition, and artificial implementation alone does not rule it out. The relevant questions concern how the system represents agents and objects, how evaluation enters, whether a qualifying relationship organizes the judgment, and which routes actually contribute. Output is evidence; it is not provenance.

10. What is fixed, and what remains open

NDT's present strength is conceptual and architectural. It proposes a formally and conceptually noncircular architecture for distinguishing moral from nonmoral normative cognition, explains how identical outputs can differ in type, integrates explicit and encoded evaluation, and makes framing and actual contribution central to judgment. Whether that boundary

can be identified through validated, operationally independent measures remains open. Its formal specification states finite consequences and exposes dependencies that can otherwise remain hidden.

The relationship roles belong to this constitutive level. Inherent valence, Relational Positioning, and perceived agency remain independently variable; access and profile-sensitive actual contribution are separate. Identity-mediated positioning, intimacy, and direct encounter are alternative routes. Shared, contrastive, and complementary-role identities are forms within the identity-mediated route. Standing-mediated positioning is retired, and strict benders are reserved for core-preserving downstream effects. No universal numerical threshold is asserted.

Several other questions remain open:

- how encoded evaluation is realized and how common it is;
- how positive and negative inherent value, each positioning route and identity form, perceived agency, functional access, strict benders, profile-sensitive contribution, and attributed recognition should be identified independently;
- whether the three positioning routes and three identity-mediated forms yield distinct transfer or process signatures and how the taxonomy should be revised if they do not;
- whether the present relationship decomposition should be subdivided, combined, or extended;
- what quantitative interaction and thresholds best model particular systems or populations;
- how actual contribution can be identified when several routes redundantly support the same output;
- how implicit subjects and collective agents can be measured;
- how frames and judgment events are individuated in empirical data;
- which challenge-handling and arbitration procedures human or artificial systems use; and
- how character assessment, expected conduct, and judgment provenance interact;
- whether any real judgment episodes instantiate the complete framed-detachment architecture rather than isolated components; and
- whether independently identified provenance groupings yield more stable intervention, transfer, process, or explanatory patterns than lexical, behavioral, character-based, or policy-based partitions.

These questions belong to four standards of assessment. **Comparative explication adequacy** asks whether NDT characterizes the target kind better than alternatives. **Formal countermodel structure** asks whether its definitions generate explicit cases in which registered conclusions fail. **Operational identification** asks whether real episodes instantiating the proposed organization can be distinguished from plausible rivals. **Calibrated measurement** requires validated indicators, parameters, uncertainty estimates, and a declared population or operating range. Success under one standard does not guarantee success under the others.

Empirical development can refine, restrict, support, or undermine applications of the theory. For example, researchers might hold words and overt choices constant while varying relationship information, compare transfer by relational similarity with transfer by rule membership, or intervene on policy and relationship routes separately. Process measures could help distinguish a requirement constructed locally by a relationship from one applied through a provenance-free rule. Artificial systems may offer tractable comparison classes because parts of their representations and policies can sometimes be manipulated more directly than human cognition.

At the model level, sustained failure to identify any complete episode under mature measures would count against full-architecture instantiation. Repeated failure of pre-specified provenance partitions to yield their registered intervention, transfer, process, or explanatory consequences would by itself count against provenance-based kind unity. Consistent outperformance by rival partitions would strengthen that evidence but is not required before it becomes adverse. Mechanistic diversity alone would not decide the latter because a cognitive kind may be multiply realized.

None of this makes falsifiability the master standard of NDT's value. Constitutive proposals are also assessed by whether they define a coherent explanandum, avoid circularity, preserve paradigm and contrast cases, integrate previously separated phenomena, and outperform rival explications. Some empirical claims—about prevalence, quantitative thresholds, psychological robustness, or implementation—are straightforwardly vulnerable to evidence. Other disputes, such as whether an allegedly agentless appraisal belongs to the moral kind, require comparative explication rather than a raw count of verbal classifications.

The distinction matters in both directions. NDT should not be weakened merely because a disputed label is applied differently, and it cannot insulate itself by redescribing every counterexample as nonmoral. Sustained failure on independently specified contrast or consequence tests can require restriction, revision, or abandonment of a proposed construct. A rival helps locate and explain that failure but is not required before evidence counts as adverse. Where adequate measurement is not yet available, that remains a registered limitation rather than support for the theory.

Conclusion: from a meaningful world to judgment

Normative Detachment Theory begins before rules are selected and before moral values are applied. Someone forming a judgment represents a world in which some distinctions already matter. A frame organizes that represented world into agents, objects, actions, relationships, possibilities, and questions. A decision model connects its evaluative structure to feasible actions, requirements, priorities, and appraisals. When connected support survives and actually contributes, a normative judgment forms.

Moral cognition is the relationally grounded subset of this process. The object receives a nonempty, noncancelling inherent-valence profile; the exact object is positioned relative to the exact subject through identity, intimacy, or direct encounter; the subject is represented with the relevant agency; and that bound relationship actually contributes through a functionally accessible process on the judgment-producing route. Downstream benders remain separately credited. This is why identical words and actions can differ in cognitive type—and why moral vocabulary, behavior, or philosophical allegiance cannot settle the question alone.

Framing ensures that the architecture does not produce one moral world viewed from nowhere. The same case can sustain several interpretations, each with its own relationships, possibilities, and locally coherent judgments. When those judgments become comparison-ready and conflict, normative polyphony appears. The result is not a theory that dissolves moral cognition into endless ambiguity. It is a theory that locates moral cognition more precisely: in the actual organization through which a meaningful, framed world becomes a normative and sometimes moral judgment.

Selected references

- Arpaly, N. (2003). *Unprincipled virtue: An inquiry into moral agency*. Oxford University Press.
- Bakhtin, M. M. (1984). *Problems of Dostoevsky's poetics* (C. Emerson, Ed. & Trans.). University of Minnesota Press. (Original work published 1929).
- Beal, B. (2018). *Raskolnikov's confession: A Dostoevskian model of moral psychology*. Doctoral dissertation, Emory University.
- Beal, B. (2020). What are the irreducible basic elements of morality? A critique of the debate over monism and pluralism in moral psychology. *Perspectives on Psychological Science*, 15, 273–290.
- Beal, B. (2021). The nonmoral conditions of moral cognition. *Philosophical Psychology*, 34, 1097–1124.
- Beal, B. (2022). The polyphony principle. *Behavioral and Brain Sciences*, 45, e222.
- Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Artificial Intelligence*, 77, 321–357.
- Herman, B. (1993). *The practice of moral judgment*. Harvard University Press.
- Horty, J. F. (2012). *Reasons as defaults*. Oxford University Press.
- Hume, D. (2000). *A treatise of human nature* (D. F. Norton & M. J. Norton, Eds.). Oxford University Press. (Original work published 1739–1740).
- Kant, I. (2012). *Groundwork of the metaphysics of morals* (M. Gregor & J. Timmermann, Eds. & Trans.). Cambridge University Press. (Original work published 1785).
- Markovits, J. (2010). Acting for the right reasons. *The Philosophical Review*, 119, 201–242.
- Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science*, 11, 481–518.
- Sliwa, P. (2016). Moral worth and moral knowledge. *Philosophy and Phenomenological Research*, 93, 393–418.
- Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge University Press.