Contextual Moral Valuation A Neuro-Cognitive Framework for Selective Moral Alignment

Momen Ghazouani
Editor-in-Chief The Ilantic Journal

April 5, 2026

Abstract

Human moral judgment operates not as a categorical evaluator of good and evil, but as a dynamic computational system integrating neural valuation signals, threat detection mechanisms, cognitive bias modulation, and psychological distance scaling. We propose the *Contextual Moral Valuation* (CMV) model, a unified theoretical framework explaining why individuals may exhibit partial sympathy toward morally transgressive agents when personal benefit is present or psychological proximity is reduced. The model posits that moral evaluation emerges from weighted interactions between reward-based neural circuits (ventral striatum), threat detection systems (amygdala), executive control regions (vmPFC), and perspective-taking networks (TPJ), with psychological distance functioning as a non-linear scaling parameter. This integration generates the *Contextual Moral Alignment Paradox* (CMAP): simultaneous condemnation and alignment with the same agent under varying contextual parameters. We formalize the model mathematically, derive testable predictions, situate it within existing theoretical frameworks, and outline future empirical directions. The CMV model offers a parsimonious explanation for selective empathy, moral disengagement, and context-dependent ethical reasoning.

1 Introduction

1.1 The Problem of Moral Inconsistency

A fundamental puzzle in moral psychology concerns the apparent inconsistency of human ethical judgment. Individuals routinely condemn harmful actions in principle while simultaneously expressing sympathy, support, or even admiration for agents who commit such actions under specific circumstances. This phenomenon transcends simple hypocrisy it represents a systematic pattern wherein the same moral agent can elicit both condemnation and approval from the same evaluator, contingent upon contextual variables.

Classical moral philosophy typically assumes moral judgments reflect stable principles or virtues (10; 1). However, empirical evidence consistently demonstrates that moral evaluations are highly context-sensitive (5; 7). People exhibit different moral standards for in-group versus out-group members (20), proximal versus distal victims (21), and personally beneficial versus costly scenarios (18).

1.2 Limitations of Existing Models

Several theoretical frameworks address aspects of moral variability:

- **Dual-Process Models** (6): Propose competition between emotion-based and reason-based systems, but do not adequately account for systematic variation in emotional responses themselves.
- **Moral Foundations Theory** (8): Identifies multiple moral dimensions but treats them as relatively stable individual differences rather than dynamically weighted computations.
- **Construal Level Theory** (21): Demonstrates that psychological distance affects abstraction but does not explicitly model moral valuation mechanisms.
- **Moral Disengagement Theory** (2): Catalogs cognitive mechanisms for justifying harmful behavior but lacks neural grounding and quantitative formalization.

No existing framework integrates neural valuation systems, cognitive biases, and psychological distance into a unified computational model of context-dependent moral judgment.

1.3 The Present Framework

We propose the *Contextual Moral Valuation* (CMV) model, which posits that moral evaluation emerges from the integration of four primary components:

1. **Reward Valuation (R)**: Neural encoding of potential benefits associated with the moral agent
2. **Threat Detection (T)**: Neural encoding of potential harms associated with the moral agent
3. **Bias Modulation (B)**: Cognitive reframing mechanisms that adjust perceived valence
4. **Psychological Distance (D)**: Spatial, temporal, social, or hypothetical separation from consequences

The model predicts that moral judgment (M) emerges from a non-linear interaction among these components, with psychological distance asymmetrically attenuating threat signals while preserving reward signals—a pattern we term the *Contextual Moral Alignment Paradox* (CMAP).

2 Theoretical Foundations

2.1 Neural Basis of Moral Valuation

2.1.1 Reward Processing: The Ventral Striatum

The ventral striatum, particularly the nucleus accumbens, encodes reward prediction and experienced value (11). Neuroimaging studies demonstrate that this region activates when

evaluating personally beneficial outcomes, even when such benefits arise from morally questionable sources (4). We hypothesize that the reward signal (R) in moral judgment reflects ventral striatal activity encoding the magnitude of personal benefit associated with a moral agent's actions.

Definition 1 (Reward Signal)**.** Let $R : \mathcal{A} \times \mathcal{C} \to \mathbb{R}^+$ be a function mapping a moral agent $a \in \mathcal{A}$ in context $c \in \mathcal{C}$ to a non-negative reward value, representing the magnitude of expected or experienced personal benefit.

2.1.2 Threat Processing: The Amygdala

The amygdala responds to threatening stimuli and plays a central role in aversive learning (14). Moral transgressions, particularly those involving harm, reliably activate the amygdala (16). We posit that the threat signal (T) represents amygdala-mediated encoding of harm magnitude and potential danger associated with the moral agent.

Definition 2 (Threat Signal)**.** Let $T : \mathcal{A} \times \mathcal{C} \to \mathbb{R}^+$ be a function mapping a moral agent a in context c to a non-negative threat value, representing the magnitude of perceived harm or danger.

2.1.3 Executive Control: The vmPFC

The ventromedial prefrontal cortex (vmPFC) integrates emotional valuation with contextual information and is crucial for value-based decision-making (3). Lesions to vmPFC result in impaired moral judgment, particularly in contexts requiring integration of competing values (12). We propose that vmPFC implements the integration function that combines R, T, and D into a unified moral evaluation.

2.1.4 Perspective-Taking: The TPJ

The temporoparietal junction (TPJ) supports mental state attribution and perspective-taking (19). This region becomes particularly active when evaluating actions based on intentions rather than outcomes (22). We suggest that TPJ activity modulates how psychological distance affects moral judgment by adjusting the weight given to distant versus proximal perspectives.

2.2 Psychological Distance as a Scaling Function

Construal Level Theory (21) demonstrates that psychological distance—encompassing spatial, temporal, social, and hypothetical dimensions—systematically affects cognitive processing. Distant events are construed more abstractly, emphasizing core features while de-emphasizing peripheral details.

Hypothesis 1 (Asymmetric Distance Scaling)**.** Psychological distance asymmetrically affects moral valuation: it attenuates perceived threat more strongly than perceived reward, leading to relatively more favorable evaluations of harmful-but-beneficial agents when they are psychologically distant.

This asymmetry may arise because:

1. **Tangibility Bias**: Concrete benefits (e.g., economic gains) remain salient across distance, whereas abstract harms (e.g., distant suffering) become less vivid.

2. **Motivated Reasoning**: Individuals are motivated to maintain positive views of beneficial agents, leading to selective attention to rewards over threats at distance.

3. **Neural Architecture**: Reward signals may be processed more independently of contextual detail than threat signals, which require specificity for adaptive responding.

2.3 Cognitive Bias Modulation

Moral judgments are subject to systematic cognitive biases that serve motivated reasoning functions (13). Key biases relevant to CMV include:

- **Self-Serving Bias**: Tendency to interpret ambiguous information in ways that serve self-interest (15)
- **Moral Disengagement**: Cognitive restructuring that allows harmful actions to be viewed as acceptable (2)
- **System Justification**: Motivation to defend and rationalize existing social arrangements (9)
- **Confirmation Bias**: Preferential processing of information consistent with existing beliefs (17)

Definition 3 (Bias Modulation Function)**.** Let $B : \mathcal{A} \times \mathcal{C} \times \mathcal{S} \to \mathbb{R}$ be a function mapping agent a, context c, and individual state $s \in \mathcal{S}$ to a bias value that amplifies or attenuates the net moral evaluation. $B > 0$ indicates biases favoring the agent; $B < 0$ indicates biases disfavoring the agent.

3 Formal Model Specification

3.1 Basic Formulation

We propose that moral judgment emerges from the following functional form:

$$M(a, c, s) = f\left(\frac{\alpha(s) \cdot R(a, c)}{D(a, c)^{\gamma}}, \frac{\beta(s) \cdot T(a, c)}{D(a, c)^{\delta}}\right) + B(a, c, s) \tag{1}$$

where:

- $M(a, c, s)$: Moral evaluation of agent a in context c by individual in state s
- $\alpha(s), \beta(s)$: Individual sensitivity parameters for reward and threat
- γ, δ: Distance scaling exponents (with $\gamma < \delta$ predicting asymmetry)
- $f(\cdot, \cdot)$: Integration function (e.g., difference, ratio, or non-linear combination)
- $D(a, c) \geq 1$: Psychological distance measure

3.2 Simplified Linear Approximation

For analytical tractability and initial empirical testing, we propose a simplified linear form:

$$M(a, c, s) = \frac{\alpha(s) \cdot R(a, c)}{D(a, c)} - \frac{\beta(s) \cdot T(a, c)}{D(a, c)} + \eta(s) \cdot B(a, c, s) \tag{2}$$

This can be rewritten as:

$$M(a, c, s) = \frac{\alpha(s) \cdot R(a, c) - \beta(s) \cdot T(a, c)}{D(a, c)} + \eta(s) \cdot B(a, c, s) \tag{3}$$

where $\eta(s)$ represents individual susceptibility to bias modulation.

3.3 The Contextual Moral Alignment Paradox

Definition 4 (CMAP)**.** The Contextual Moral Alignment Paradox occurs when an individual simultaneously:

1. Acknowledges that agent a has committed morally transgressive actions ($T(a, c) > \theta_T$ for some threshold θ_T)
2. Expresses positive moral evaluation or sympathy toward a ($M(a, c, s) > 0$)

Proposition 1 (Conditions for CMAP)**.** CMAP emerges when:

$$R(a, c) > \frac{\beta(s)}{\alpha(s)} \cdot T(a, c) - \frac{D(a, c)}{\alpha(s)} \cdot \eta(s) \cdot B(a, c, s) \tag{4}$$

This condition is most likely satisfied when:

- $R(a, c)$ is high (substantial personal benefit)
- $D(a, c)$ is large (consequences are psychologically distant)
- $B(a, c, s) > 0$ (cognitive biases favor the agent)
- $\beta(s)/\alpha(s)$ is low (individual weighs rewards more than threats)

3.4 Individual Differences

The model incorporates individual variability through personalized parameters:

- α_i, β_i: Individual reward and threat sensitivity (may correlate with personality traits such as psychopathy, empathy, or utilitarianism)
- η_i: Individual susceptibility to motivated reasoning
- Cultural factors may shift baseline values of α, β, and the functional form of D (e.g., collectivist vs. individualist distance metrics)

4 Integration with Existing Theories

4.1 Dual-Process Theory

Greene's dual-process model (6) proposes that moral judgment results from competition between automatic emotional responses and controlled cognitive processes. The CMV model extends this by:

1. Specifying what drives emotional responses (R and T as competing valuation signals)
2. Modeling how cognitive processes modulate these signals (B as bias-driven reinterpretation)
3. Adding psychological distance as a factor that shifts the balance between emotion and cognition

4.2 Construal Level Theory

CLT (21) predicts that psychological distance leads to more abstract, value-focused processing. CMV incorporates this through the distance parameter D but makes a specific prediction: distance asymmetrically preserves abstract benefits (rewards) while diminishing concrete harms (threats).

4.3 Moral Foundations Theory

MFT (8) identifies multiple moral domains (harm/care, fairness, loyalty, authority, purity). CMV does not contradict MFT but operates at a different level of analysis: while MFT describes *what* moral values people hold, CMV models *how* contextual factors dynamically weight those values in producing judgments.

4.4 Social Identity Theory

Social identity processes (20) affect D (in-group members are psychologically closer) and B (group-serving biases). The CMV model naturally accommodates in-group favoritism: in-group harmful agents have lower D but benefit from positive B, while out-group harmful agents have higher D and negative B.

5 Theoretical Predictions

5.1 Main Predictions

Hypothesis 2 (Distance-Benefit Interaction)**.** For agents providing personal benefit ($R > 0$) while causing harm ($T > 0$), moral evaluation M increases with psychological distance D.

Hypothesis 3 (Asymmetric Distance Scaling)**.** The attenuation effect of distance is stronger for threat signals than reward signals: $\gamma < \delta$ in the general formulation.

Hypothesis 4 (Bias Amplification Under Uncertainty)**.** When information about the agent's actions is ambiguous, bias modulation B has a stronger effect on moral judgment.

Hypothesis 5 (Neural Dissociation)**.** Ventral striatal activation correlates with R and remains relatively stable across distance, while amygdala activation correlates with T and decreases more sharply with distance.

5.2 Boundary Conditions

The model predicts weakened or reversed effects under certain conditions:

- **Direct Victimization**: When the evaluator is a direct victim, D approaches its minimum, eliminating distance-based attenuation
- **Extreme Transgressions**: For severe moral violations (e.g., genocide), T may remain high even at distance due to salient categorical information
- **Deontological Commitments**: Individuals with strong rule-based moral convictions may show reduced influence of R and B
- **Reputation Concerns**: Public evaluation contexts may reduce B due to social desirability pressures

6 Implications and Applications

6.1 Political Psychology

The CMV model explains why electorates sometimes support leaders who commit ethical violations:

- Economic prosperity (R) can outweigh distant human rights abuses (T at high D)
- Partisan identity strengthens B through motivated reasoning
- Media framing can manipulate perceived D (emphasizing local benefits, de-emphasizing distant harms)

6.2 Organizational Ethics

In corporate contexts, the model predicts:

- Employees may tolerate unethical practices when they personally benefit (high R) and victims are distant (high D)
- Whistleblowing likelihood inversely correlates with personal benefit and directly correlates with proximity to victims
- Organizational culture can shift α, β, and η through normative influence

6.3 Moral AI Development

For AI systems designed to model human moral reasoning:

- Simple rule-based or pure utilitarian systems fail to capture context-dependent moral evaluation
- Realistic moral AI must incorporate distance-scaled valuation and bias parameters
- Understanding CMV mechanisms can help design AI systems that are robust to manipulation through framing

6.4 Intervention Design

The model suggests intervention strategies to reduce moral disengagement:

- **Psychological Proximity Induction**: Vivid narratives, personal testimonies, or virtual reality experiences can reduce D
- **Bias Awareness Training**: Metacognitive interventions can reduce η by increasing awareness of motivated reasoning
- **Reward Reframing**: Highlighting long-term costs of short-term benefits can adjust perceived R

7 Empirical Research Directions

7.1 Behavioral Studies

Proposed Paradigm: Vignette-based studies manipulating:

- Agent benefit provision (high vs. low R)
- Agent harm commission (high vs. low T)
- Psychological distance (local vs. remote victims)
- Measurement: Moral evaluation scales, sympathy ratings, behavioral intentions

Key Comparisons:

1. $R^+T^+D_{high}$ vs. $R^+T^+D_{low}$: Tests distance effect on CMAP
2. $R^+T^+D_{high}$ vs. $R^-T^+D_{high}$: Tests reward necessity for CMAP
3. Individual differences in α, β, η using personality and value scales

7.2 Neuroscientific Studies

fMRI Predictions:

- Ventral striatum activity should correlate with R across distance conditions
- Amygdala activity should correlate with T but show distance-dependent attenuation
- vmPFC should show integration patterns consistent with the model's functional form
- TPJ activity should modulate with psychological distance manipulations

TMS/Lesion Studies: Disruption of vmPFC should impair integration of R, T, and D, leading to more inconsistent moral judgments.

7.3 Computational Modeling

Agent-based simulations can explore:

- Population-level moral dynamics when individuals follow CMV
- Emergence of collective support for harmful-but-beneficial leaders
- Effects of media manipulation of perceived D
- Cultural evolution of α, β, and η distributions

8 Limitations and Future Directions

8.1 Current Limitations

1. **Simplified Functional Form**: The linear approximation may not capture true neural integration dynamics
2. **Static Parameters**: In reality, α, β, and η may vary dynamically based on context
3. **Limited Scope**: The model focuses on benefit-harm trade-offs but does not fully address other moral dimensions (e.g., fairness, purity)
4. **Measurement Challenges**: Operationalizing R, T, D, and B in real-world contexts requires validated instruments

8.2 Extensions

Future theoretical work should address:

- **Non-linear Interactions**: Developing more sophisticated integration functions (e.g., multiplicative, threshold-based)
- **Temporal Dynamics**: Modeling how moral evaluations change over time as information accumulates

- **Multi-Agent Contexts**: Extending to scenarios involving multiple moral agents with interdependent actions
- **Cultural Variation**: Systematically mapping cultural differences in parameter distributions
- **Developmental Trajectory**: Modeling how CMV parameters change across the lifespan

9 Conclusion

The Contextual Moral Valuation (CMV) model provides a unified theoretical framework for understanding context-dependent moral judgment. By integrating neural valuation systems (reward and threat processing), cognitive bias mechanisms, and psychological distance, the model explains the Contextual Moral Alignment Paradox: why individuals can simultaneously condemn and sympathize with the same moral agent.

The model makes specific, testable predictions about behavioral patterns, neural activation, and individual differences. It integrates insights from dual-process theory, construal level theory, moral foundations theory, and social identity theory while adding novel mechanistic specificity. The CMV framework has significant implications for understanding political support, organizational ethics, moral AI development, and intervention design.

By formalizing the computational mechanisms underlying selective moral alignment, the CMV model contributes to a more realistic, neurally grounded, and predictive science of moral psychology. Future empirical work testing the model's predictions will refine our understanding of when, why, and how humans deviate from principled moral consistency—not as a failure of reasoning, but as a predictable consequence of the architecture of human moral cognition.

Keywords: moral judgment, psychological distance, reward processing, threat detection, cognitive bias, dual-process theory, neuroethics, contextual moral alignment paradox

References

[1] Aristotle. (350 BCE). *Nicomachean Ethics*.

[2] Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193-209.

[3] Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex*, 10(3), 295-307.

[4] Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., & Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, 84(6), 3072-3077.

[5] Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105-2108.

[6] Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389-400.

[7] Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814-834.

[8] Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

[9] Jost, J. T., & Banaji, M. R. (2004). A decade of system justification theory: Accumulated evidence of conscious and unconscious bolstering of the status quo. *Political Psychology*, 25(6), 881-919.

[10] Kant, I. (1785). *Groundwork of the metaphysics of morals*.

[11] Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*, 21(16), RC159.

[12] Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138), 908-911.

[13] Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480-498.

[14] LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, 23(1), 155-184.

[15] Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, 130(5), 711-747.

[16] Moll, J., de Oliveira-Souza, R., Eslinger, P. J., Bramati, I. E., Mourão-Miranda, J., Andreiuolo, P. A., & Pessoa, L. (2002). The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *Journal of Neuroscience*, 22(7), 2730-2736.

[17] Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175-220.

[18] Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes*, 109(2), 134-141.

[19] Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, 19(4), 1835-1842.

[20] Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations*, 33(47), 74.

[21] Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, 117(2), 440-463.

[22] Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, 104(20), 8235-8240.