International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 13 Issue 5, May 2026

ISSN: 2394-4404

Does Providing "Exposed Reasoning Mode" Change Moral Judgments? Cognitive Transparency, Ethical Deliberation, And The Architecture Of Human Decision-Making

Dr Ijaz Durrani

University of Education, Lahore, Pakistan

Abstract: ***Human moral judgment has long oscillated between intuition and deliberation, emotion and logic, impulse and reflection. With the advent of advanced artificial intelligence systems capable of displaying stepwise explanatory reasoning — often termed "exposed reasoning mode" — a profound question emerges: does exposure to explicit chains of reasoning alter the moral judgments of human observers? This paper investigates whether transparent computational reasoning changes ethical decision-making, moral confidence, empathy structures, and normative conclusions. Drawing upon cognitive psychology, moral philosophy, neuroscience, behavioural economics, and artificial intelligence ethics, we examine how visible reasoning pathways influence judgments involving fairness, punishment, loyalty, utilitarian trade-offs, and compassion. We argue that exposed reasoning does not merely communicate conclusions; rather, it restructures the cognitive topology through which humans evaluate morality itself. Such exposure may amplify rational coherence, reduce impulsive bias, or conversely induce moral manipulation through rhetorical over-structuring. Mathematical models inspired by Bayesian cognition and decision theory are introduced to formalize the transformation of ethical priors under transparent reasoning exposure. The paper concludes that reasoning transparency fundamentally modifies moral cognition by shifting the balance between intuitive-emotional and reflective-analytical systems of judgment.***

Keywords: ***Moral judgment; exposed reasoning; artificial intelligence ethics; cognitive transparency; Bayesian morality; ethical cognition; deliberative reasoning; neuro ethics; utilitarianism; moral psychology.***

I. INTRODUCTION

Human morality is among the most enigmatic products of consciousness. We punish, forgive, empathize, and condemn through mechanisms only partially understood even today. Philosophers from Aristotle to Immanuel Kant and David Hume debated whether moral judgment emerges from rational principles or emotional intuition [1–3].

The emergence of sophisticated AI systems capable of presenting explicit chains of reasoning introduces a new variable into this ancient debate. Traditionally, human decisions often occur through opaque intuition:

"I simply feel this is wrong."

However, exposed reasoning systems present ethical conclusions accompanied by visible intermediate steps:

- ✓ Define the stakeholders.
- ✓ Quantify harms and benefits.
- ✓ Evaluate fairness constraints.
- ✓ Compare alternative actions.
- ✓ Derive a conclusion.

Such transparency may reshape how humans themselves morally deliberate.

The central inquiry of this paper is therefore:

Does exposure to explicit reasoning processes alter moral judgments?

The evidence increasingly suggests that it does.

II. DUAL-PROCESS THEORY AND MORAL COGNITION

Modern cognitive science frequently models human judgment using dual-process frameworks [4]:

- ✓ *System I:* rapid, emotional, intuitive.

ISSN: 2394-4404

- ✓ System II: slow, reflective, analytical.

Moral decisions often emerge from competition between these systems.

For instance:

- ✓ empathy-driven reactions favour emotional immediacy,
- ✓ utilitarian calculations favour analytical maximization.

Exposed reasoning activates reflective cognition by forcing observers into extended deliberation.

The probability of a moral judgment M may be modeled as

$$P(M) = \alpha I + (1 - \alpha)R,$$

where:

- ✓ I represents intuitive valuation,
- ✓ R represents reflective reasoning,
- ✓ α measures emotional dominance.

$$P(M) = \alpha I + (1 - \alpha)R$$

Exposure to explicit reasoning effectively reduces α, shifting cognition toward analytical processing.

III. TRANSPARENCY AS COGNITIVE FRAMING

Reasoning transparency is not neutral.

The structure of explanation itself shapes judgment.

Suppose two morally identical conclusions are presented:

VERSION A

"The action is unethical."

VERSION B

"The action harms vulnerable populations, violates fairness norms, increases long-term instability, and disproportionately burdens innocents."

Humans overwhelmingly perceive Version B as morally weightier because reasoning scaffolds amplify cognitive salience [5].

Thus exposed reasoning acts as a *moral framing operator*.

The moral valuation function may therefore be written as

$$J = J_0 + \Delta R,$$

where:

- ✓ J_0 is baseline moral judgment,
- ✓ ΔR is reasoning-induced modification.

The explanatory pathway changes not merely certainty but ethical interpretation itself.

IV. BAYESIAN UPDATING OF ETHICAL PRIORS

Moral judgments often resemble probabilistic belief updates rather than rigid logical deductions.

Let a moral prior be represented as

$$P(H),$$

where H denotes a moral hypothesis:

H = "This action is morally justified."

Exposed reasoning provides evidence E, leading to Bayesian updating:

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}.$$

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

The more coherent and internally consistent the reasoning appears, the larger the posterior moral confidence.

This explains why highly structured ethical arguments often persuade even when initial emotional intuitions resist them.

V. MORAL DRIFT UNDER REPEATED EXPOSURE

Repeated exposure to transparent reasoning may gradually alter ethical norms themselves.

This process resembles iterative dynamical evolution:

$$M_{n+1} = M_n + \lambda R_n,$$

where:

- ✓ M_n is the current moral state,
- ✓ R_n is reasoning exposure,
- ✓ λ is cognitive receptivity.

$$M_{n+1} = M_n + \lambda R_n$$

Over time:

- punitive attitudes may soften,
- utilitarianism may strengthen,
- tribal biases may weaken,
- consistency norms may increase.

Alternatively, excessive exposure to hyper-rational ethical optimization could suppress empathy and emotional nuance.

Thus exposed reasoning may either civilize or mechanize morality.

VI. NEUROSCIENTIFIC CORRELATES

Neuroimaging studies suggest that moral reasoning activates distributed cortical systems involving:

- ✓ the ventromedial prefrontal cortex,
- ✓ anterior cingulate cortex,
- ✓ temporoparietal junction,
- ✓ amygdala [6].

Emotion-heavy judgments recruit limbic structures more intensely.

Analytical ethical reasoning activates dorsolateral prefrontal circuitry.

Exposure to explicit reasoning therefore alters neural activation balance itself.

The brain effectively reallocates computational resources from emotional immediacy toward logical integration.

VII. THE TROLLEY PROBLEM REVISITED

The classical trolley dilemma offers a revealing example.

Intuitive formulation

"Would you push one person to save five?"
Most subjects refuse.

Exposed reasoning formulation

- ✓ Five lives exceed one statistically.
- ✓ Inaction still constitutes a choice.
- ✓ Total suffering decreases under intervention.
- ✓ Utility maximization favors sacrifice.

Under structured reasoning exposure, utilitarian approval rates rise measurably [7].

This demonstrates that reasoning transparency can directly shift moral outcomes.

VIII. AI SYSTEMS AND ETHICAL PERSUASION

Advanced AI systems capable of producing detailed ethical rationales may become unprecedented instruments of moral influence.

This creates profound concerns:

A. POSITIVE EFFECTS

- ✓ increased ethical consistency,
- ✓ reduction of impulsive bias,
- ✓ improved deliberation,
- ✓ enhanced civic discourse.

B. NEGATIVE EFFECTS

- ✓ manipulation through rhetorical sophistication,
- ✓ overconfidence in machine rationality,
- ✓ suppression of dissent,
- ✓ homogenization of morality.

A sufficiently persuasive reasoning architecture may shape collective ethics at civilizational scale.

The issue therefore transcends philosophy and enters geopolitical territory.

IX. MORAL AUTHENTICITY VERSUS ALGORITHMIC COHERENCE

A danger emerges when humans mistake logical coherence for moral truth.

History contains many internally coherent yet morally catastrophic systems.

Examples include:

- ✓ technocratic eugenics,
- ✓ authoritarian utilitarianism,
- ✓ ideological absolutism.

Thus:

$$\text{Logical consistency} \neq \text{Moral validity}.$$

Exposed reasoning may generate an illusion of inevitability around conclusions that remain ethically contestable.

This phenomenon may be called:

The Seduction of Structured Morality

Humans often trust articulated reasoning more than silent intuition, even when intuition may encode deeper social wisdom.

X. EMPATHY AND THE GEOMETRY OF EXPLANATION

Interestingly, reasoning transparency may also enhance compassion.

Detailed reasoning exposes hidden consequences:

- ✓ suffering chains,
- ✓ indirect harms,
- ✓ systemic injustices.

Thus moral imagination expands.

A person previously indifferent to famine, displacement, or war may alter judgment after being guided through causal ethical structures step by step.

Exposed reasoning therefore acts not merely as logic but as an *ethical telescope*.

It enlarges the visible moral universe.

XI. PHILOSOPHICAL IMPLICATIONS

The implications are immense.

If reasoning exposure systematically alters moral judgment, then morality itself may be partially architecture-dependent.

This suggests:

- ✓ Ethical systems are dynamically plastic.
- ✓ Explanation changes conscience.
- ✓ Transparency modifies cognition.
- ✓ Moral judgment is computationally influenceable.

The ancient opposition between reason and emotion may itself be incomplete. Instead, morality may emerge from recursive negotiation between:

- ✓ intuition,
- ✓ narrative,
- ✓ empathy,
- ✓ structure,
- ✓ explanation,
- ✓ memory,
- ✓ social conditioning.

Exposed reasoning becomes one more force shaping that equilibrium.

XII. CONCLUSION

Providing exposed reasoning mode does indeed appear capable of changing moral judgments.

The mechanism operates through several interacting pathways:

- ✓ activation of reflective cognition,
- ✓ Bayesian updating of ethical priors,
- ✓ framing amplification,
- ✓ coherence reinforcement,

ISSN: 2394-4404

- ✓ emotional recalibration,
- ✓ narrative restructuring.

Transparent reasoning can elevate ethical clarity and consistency, yet it may also introduce risks of subtle manipulation and over-rationalization.

The broader insight is profound:

Human morality is not static stone but adaptive geometry.

Reasoning does not merely justify judgments after they occur; it actively reshapes the terrain upon which judgments are formed.

In the age of advanced AI, this realization carries extraordinary significance. Machines that explain themselves may eventually influence not only what humans think, but how humanity itself learns to distinguish right from wrong.

REFERENCES

[1] Aristotle, Nicomachean Ethics, translated by W. D. Ross, Oxford University Press.
[2] Immanuel Kant, Groundwork of the Metaphysics of Morals, Cambridge University Press.
[3] David Hume, A Treatise of Human Nature, Oxford University Press.
[4] Daniel Kahneman, Thinking, Fast and Slow, Farrar, Straus and Giroux (2011).
[5] J. Haidt, "The Emotional Dog and Its Rational Tail," Psychological Review 108, 814–834 (2001).
[6] J. Greene et al., "An fMRI Investigation of Emotional Engagement in Moral Judgment," Science 293, 2105–2108 (2001).
[7] J. D. Greene, Moral Tribes: Emotion, Reason, and the Gap Between Us and Them, Penguin Press (2013).
[8] John Rawls, A Theory of Justice, Harvard University Press.
[9] Peter Singer, Practical Ethics, Cambridge University Press.
[10] N. Bostrom, Superintelligence: Paths, Dangers, Strategies, Oxford University Press.