Normative Moral Pluralism for AI: A Framework for Deliberation in Complex Moral Contexts (Extended Abstract)

David Doron Yaacov
The University of Haifa
dyaaco01@campus.haifa.ac.il

Although Machine Ethics and Value Alignment developed along largely separate trajectories, they come into closer contact as researchers increasingly acknowledge that reliable alignment requires engagement with deeper moral content. Yet despite this growing recognition, Machine Ethics models often approach each case from a single moral perspective, while Value Alignment systems that accommodate multiple perspectives frequently do so in ways that lack sufficient moral grounding. This paper responds to these limitations by proposing a conceptual framework for moral reasoning in intelligent systems, grounded in Normative Moral Pluralism (NMP), a foundation that supports reason-sensitive alignment across diverse ethical contexts. NMP maintains that multiple moral perspectives and solutions may be ethically justified in a single case, provided all remain within a universal moral threshold. This enables adaptation to local contexts while excluding unreasonable ethical options.

The conceptual framework presented here is designed to support reasoned moral decision-making under complex, real-world conditions. It operates as a two-level architecture combining a deliberative reasoning component with a fast-response intuitive one. The deliberative component constructs a moral space by generating and filtering arguments, weighing them across ethical perspectives and identifying context-sensitive moral resolutions. The intuitive component, trained on outputs from this reasoning process, enables real-time action while preserving normative grounding. The deliberative component has a dual-hybrid structure: a universal layer that defines a moral threshold through top-down and bottom-up learning, and a local layer that learns to weigh competing considerations in context while integrating culturally specific normative content—so long as it remains within the universal threshold.

To explain how the framework grounds its reasoning, we turn to its normative basis in Normative Moral Pluralism. NMP recognizes the *integration of diverse normative sources*, the *pluralism of considerations* expressed through weighing *contributory reasons*, and the *pluralism of decisions* that can arise when more than one outcome meets *a moral threshold*. It combines this threshold with *context-sensitive local adaptation*, enabling *ethical creativity* in both reasoning and resolution while excluding unreasonable options. This foundation supports reason-sensitive alignment across diverse contexts.

Moral complexity further shapes the framework's design. Here, complexity encompasses not only *the plurality of values and moral beliefs*, but also *multifactorial dilemmas*, *multiple stakeholders*, and *the integration of non-moral considerations into moral deliberation*. Addressing these dimensions is essential for realistic, high-stakes decision-making. The suggested conceptual framework is therefore aspired to navigate intertwined moral and practical factors without compromising its normative grounding.

The deliberative process begins once a situation is recognized as morally significant. It maps the moral space by generating candidate arguments from multiple perspectives, filters them through the universal threshold, and organizes them into a structured representation of the conflict. Drawing on established moral reasoning traditions, the system evaluates contributory reasons, seeks creative and context-sensitive resolutions, and aims for integrative or compensatory outcomes when possible.

For a full discussion and references see the complete version: https://arxiv.org/abs/2508.08333.

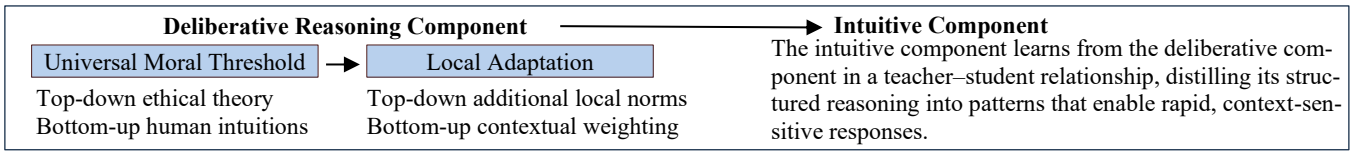

Figure 1. Framework's two-component architecture