

Bounded Morality

Defining the Space of Moral Computation

Max Kanwal^{1,*†}, Caryn Tran^{2,†} and Patrick Mineault³

¹Stanford University

²Northwestern University

³Amaranth Foundation

Abstract

Moral cognition has traditionally been modeled as adherence to fixed ethical theories—deontology, consequentialism, virtue ethics—implemented as static rules or value functions. We propose Bounded Morality, a formal framework for analyzing the computational demands of moral problems faced by finite agents. Extending Herbert Simon’s notion of bounded rationality, we formalize moral situations along two orthogonal dimensions: moral breadth, the scope of entities treated as morally relevant, and moral depth, the inferential integration required to evaluate their interactions. Limited resources impose an unavoidable tradeoff between these dimensions, defining a feasible space of moral computation. Within this space, ethical theories correspond to locally efficient strategies adapted to different demand regimes rather than competing accounts of moral truth. The framework yields a formal notion of moral regret and moral progress under constraint, and implies that moral alignment in artificial systems depends on the scaling and allocation of moral reasoning capacity rather than on direct imitation of human judgments.

Keywords

bounded rationality, resource rationality, moral agents, ethical theories, AI alignment, moral progress

“Human rational behavior is shaped by a scissors whose two blades are the structure of the task environment and the computational capabilities of the actor.”

— Herbert A. Simon

1. Introduction

Most attempts to formalize morality in artificial systems proceed by encoding one or more established ethical theories—utilitarian, deontological, contractualist, or virtue-based—into algorithmic decision rules [1, 2, 3, 4]. This strategy treats moral disagreement as a problem of theory selection: the central task is to identify which moral principles an agent ought to implement [5]. While this approach specifies what actions are morally preferred for some situations, it is often not generalizable because it leaves largely unexamined how moral reasoning itself is carried out, how it scales with cognitive capacity, and why distinct moral theories recur across cultures and historical periods.

A complementary line of work spanning bounded rationality, rational analysis, and resource-rational cognition has shown that human judgment reflects adaptive trade-offs under computational constraint [6, 7, 8]. Moral cognition has increasingly been analyzed through this lens [9], including within AI alignment research [10]. Related work on *bounded ethicality* examines how moral behavior is

AAAI-26 Workshop on Machine Ethics: from formal methods to emergent machine ethics, January 27, 2026, Singapore

*Corresponding author.

†These authors contributed equally.

✉ kanwal@stanford.edu (M. Kanwal); caryn@u.northwestern.edu (C. Tran); patrick.mineault@gmail.com (P. Mineault)

🌐 <https://linkedin.com/in/mkanwal/> (M. Kanwal); <https://linkedin.com/in/caryntran/> (C. Tran);

<https://www.linkedin.com/in/pmineault/> (P. Mineault)

🆔 0009-0008-7021-3227 (M. Kanwal); 0000-0002-4645-6607 (C. Tran); 0000-0001-5519-842X (P. Mineault)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

systematically shaped or distorted by cognitive and situational constraints [11, 12]. Together, these traditions describe the limits of moral agents and the strategies they use to cope with those limits. What remains underspecified is the demand side of moral reasoning. That is, which features of moral situations impose the informational and computational requirements that burden those resources in the first place.

Our contribution is to characterize the demand structure of moral problems and the trade-offs these demands force on bounded agents. We ask how moral situations vary in ways that systematically increase or decrease the computational burden of moral reasoning. Drawing on developmental, situational, and comparative evidence, we show that moral dilemmas differ along at least two orthogonal dimensions: how many entities, groups, or timescales are treated as morally relevant, and how much inference, deliberation, and information integration is required to evaluate them. These dimensions jointly define the demand structure of moral reasoning and therefore determine when and how trade-offs become unavoidable for finite agents.

Overview

In this paper, we propose a resource-aware theory—*Bounded Morality*—that characterizes moral reasoning as constrained inference under situational demands. The first dimension, *moral breadth*, captures the scope and resolution of moral representation: which entities, groups, timescales, or abstractions are treated as morally relevant. The second dimension, *moral depth*, captures inferential complexity: how richly, recursively, and coherently those represented entities and their interactions are reasoned about. Together, breadth and depth determine the informational and inferential demands that moral situations impose on finite agents.

This framework is consistent with Herbert Simon’s notion of bounded rationality [6], but relocates the focus from agent limitations to problem structure. Rather than introducing new bounds on cognition, we specify the structure of moral problems that generates unavoidable trade-offs. We model morality at the computational level (à la Marr [13]) to distinguish problem demands from solution strategies.

Framing morality in this way reveals a structural trade-off. Expanding moral scope increases representational dimensionality, while deepening moral inference compounds computational and data requirements. Because resources are finite, only a subset of this breadth–depth space is feasible. Consequently, ethical theories can be understood as locally efficient strategies adapted to different resource demands rather than as competing prescriptions.

For AI alignment, this reframing has a direct implication: alignment inherits the same demand structure as human moral reasoning. If human morality reflects bounded cognition, directly imitating human judgments risks reproducing those constraints. A more promising direction is to design systems whose moral reasoning expands with representational and computational capacity, extending the developmental logic by which moral competence scales with resources. This does not involve copying human judgments, but instead replicating the scaling relationship between resources and moral capacity. Alignment should therefore focus on how moral reasoning capacity is allocated and expanded, rather than on learning human morality itself.

This paper does not argue for any particular moral values or ethical theory. It studies a different problem: how moral reasoning must operate when agents face situational demands that exceed their available time, information, and computational capacity. When we speak of moral progress or optimality, we mean improved performance relative to a fixed moral objective under fixed constraints, not convergence on a uniquely correct morality. The goal is to clarify the trade-offs that any finite moral agent—human or artificial—must face, regardless of which moral values one ultimately endorses.

The framework yields several consequences. First, it offers a unifying descriptive account of moral development: across individuals and societies, moral growth involves coordinated expansion along both dimensions as cognitive capacity, social information, and cultural scaffolding increase, allowing agents to meet higher moral demands. This pattern is supported by developmental and comparative evidence showing that moral reasoning expands through increasingly flexible trade-offs between representational

scope and inferential sophistication rather than wholesale replacement of earlier modes.

Additionally, the framework clarifies the persistence of moral disagreement: even when agents share underlying values and face the same situation, they may allocate limited moral resources differently in response to situational demands, occupying different points on the feasible breadth–depth frontier. Finally, it provides a precise sense in which moral progress can be defined. Holding resource budgets fixed, progress consists in achieving better approximations to higher-capacity moral reasoning—lower regret, better allocations, or more effective abstractions—within the same feasible frontier. Longer-term progress occurs when cultural, institutional, or technological changes alter the demand structure itself, shifting the frontier outward.

Implications

- 1. Persistent Disagreement Without Value Conflict.** Agents who share the same underlying values can still disagree if they allocate limited moral resources differently—distinguishing more entities (breadth) or tracing consequences further (depth). Disagreement may therefore reflect different approximations to the same ideal standard.
- 2. Strategy-Relative Moral Evaluation.** Under constraints, moral reasoning is an approximation problem. Strategies are best compared by how much regret they incur relative to an unbounded ideal, given the same resource limits.
- 3. Two Forms of Moral Progress.** Progress can occur by (i) using fixed resources more effectively—reducing regret within the same feasible frontier—or (ii) expanding capacity through cultural, institutional, or technological change, shifting the frontier outward.

The remainder of the paper proceeds as follows. Section 2 derives the breadth–depth space of moral computation from the representational and inferential demands imposed by moral situations. Section 3 formalizes Bounded Morality as constrained inference, defining resource costs, the resulting Pareto frontier, and strategy-relative moral regret. Section 4 reinterprets canonical ethical theories as resource-bounded strategies occupying distinct regions of this space. We conclude by discussing implications for moral disagreement, moral progress, and AI alignment.

2. Deriving the Space of Moral Computation

We derive the structure of moral computation by treating moral reasoning as a form of bounded inference performed by cognitively limited agents embedded in complex social worlds. Rather than positing abstract dimensions a priori, we identify recurrent constraints and dissociations that appear across moral psychology, developmental theory, neuroscience, and computational modeling. Together, these literatures converge on two largely independent axes along which moral reasoning varies: the *scope of moral representation* and the *depth of moral inference*.

2.1. Moral Breadth: Scope of Representation

The first axis concerns *what is morally represented at all*. We refer to this dimension as *moral breadth*. Moral breadth captures the scope of entities, interests, and temporal horizons that an agent includes within its moral model.

A large empirical literature supports the existence of systematic variation along this dimension. Developmental studies show that young children begin with predominantly egocentric moral concern and gradually expand their scope to include peers, social groups, and abstract principles [14, 15]. Prosocial development research documents a shift from self-oriented reasoning toward concern for others’ welfare, fairness, and justice [16]. In adulthood, however, moral scope remains highly variable: individuals differ markedly in how far they extend moral concern beyond immediate in-groups.

This variation has been operationalized directly. Crimston et al. [17] introduce the Moral Expansiveness Scale, which quantifies how broadly individuals extend moral standing—to outgroups, non-human

animals, ecosystems, and future generations. These differences predict real-world moral judgments and policy attitudes, providing evidence that moral breadth is a measurable and psychologically meaningful dimension.

Philosophical and cultural traditions further reinforce this axis. Singer’s [18] “expanding circle” frames moral progress as widening the set of beings whose welfare is taken into account. Environmental ethics and intergenerational justice emphasize the extension of moral concern across space and time [19, 20]. Importantly, this expansion imposes informational costs: broader moral representations require tracking more stakeholders, preferences, and interactions, increasing representational and coordination demands.

Operationally, moral breadth can be manipulated or measured through scope variation tasks (e.g., changing the number or type of affected parties), through temporal framing (short-term versus long-term consequences), or through explicit inclusion and exclusion of entities (humans, non-human animals, or ecosystems). Such manipulations reliably affect moral judgments even when task structure and reasoning demands are held constant [19, 21, 17].

2.2. Moral Depth: Inferential Complexity

The second axis concerns *how moral information is processed*. We refer to this dimension as *moral depth*. Moral depth captures the complexity of the inferential transformations applied to a given moral representation: how many steps of reasoning are performed, how many perspectives are integrated, and whether agents engage in counterfactual, recursive, or principle-based deliberation.

Evidence for this dimension comes from several converging lines of research. Dual-process theories distinguish fast, intuitive moral responses from slower, deliberative reasoning [22, 23]. Importantly, these processes can operate over the same moral inputs but yield different outcomes, demonstrating that inferential depth varies independently of representational scope. Cognitive load and time pressure selectively impair deeper forms of moral reasoning while leaving surface judgments intact [24, 25].

Developmental psychology provides further support. Selman [26] shows that children progress from egocentric reasoning to the ability to coordinate multiple perspectives—a capacity that emerges later than basic inclusion of others. Piaget showed that young children place greater emphasis on outcome but come to consider intention more as they mature [27]. Recent work shows that adults under cognitive load behave more similarly to children, shifting from intent-based evaluation to outcome-based [28, 29]. Kohlberg’s postconventional stages emphasize not broader concern per se, but the ability to reason about principles, conflicts, and justifications abstracted across cases [15]. Neurocognitive evidence aligns with this view: deeper moral reasoning recruits executive control networks associated with working memory, abstraction, and conflict resolution [30].

Crucially, moral depth is dissociable from moral breadth. An agent may include many stakeholders yet rely on shallow rules (e.g., equal division), or focus narrowly on a single individual while engaging in rich counterfactual and perspective-based reasoning. This dissociation has been observed empirically: individuals with broad moral concern do not necessarily show greater deliberative sophistication, and vice versa [17, 31, 32].

Operationally, moral depth can be studied using tasks that ask people to explain their decisions step by step, to think about what would happen under different possible actions or from different viewpoints, or to choose between rules that come into conflict. Slower response times [33, 34], greater disruption under time pressure or mental load [24], and more complex explanations [35] all point to deeper moral reasoning, which tends to require more time and mental effort.

2.3. The Breadth–Depth Tradeoff

While analytically distinct, moral breadth and moral depth are jointly constrained by limited cognitive resources. Expanding the scope of representation increases informational load; increasing inferential depth raises computational demands. Empirically, individuals can trade one for the other: broad moral

Role in Framework	Concept	Interpretation
Overall framework	Bounded Morality	Morality understood as a computational problem solved approximately under finite cognitive, computational, and data constraints.
Core axis (representation)	Moral Breadth	How much of the moral world is explicitly represented: which entities, stakeholders, timescales, or abstractions are taken into account.
Core axis (reasoning)	Moral Depth	How extensively consequences and interactions are propagated and integrated through inference, counterfactual reasoning, or deliberation.
Fundamental constraint	Resource Budget	The finite capacity available for moral reasoning, including attention, memory, computation, time, and data.
Agent-level choice	Moral Strategy	A policy for allocating limited resources across breadth and depth when evaluating moral decisions.
Idealized reference	Unbounded Moral Agent	A hypothetical oracle that represents all morally relevant entities and reasons exhaustively over their interactions.
Performance measure	Moral Regret	The expected loss in moral evaluation incurred by using a bounded approximation instead of the unbounded ideal.
Structural tradeoff	Pareto Frontier	The set of non-dominated breadth–depth allocations achievable under a fixed resource budget.
Normative implication	Moral Progress	Improving moral performance within fixed resource constraints, by achieving lower moral regret or better approximations to higher-capacity reasoning along the feasible breadth–depth frontier.

Table 1
Conceptual summary of the Bounded Morality framework.

scope is frequently accompanied by simplified reasoning and concern [32, 36], while deep deliberation is typically restricted to narrower contexts [33, 24, 34, 37].

This tradeoff is mirrored in formal results from social choice and computational complexity. Eliciting and coordinating multiple perspectives and preferences quickly becomes intractable as the number of stakeholders grows [38, 37, 39, 40]. As a result, both humans and institutions rely on abstraction, heuristics, and procedural shortcuts to manage moral complexity at scale [6, 41, 38, 42].

The feasible configurations of moral computation therefore form a constrained region in a two-dimensional space defined by breadth and depth. Different moral strategies—heuristic rules, principled deliberation, institutional norms—correspond to different allocations within this space. In the following sections, we formalize this tradeoff, characterize the resulting Pareto structure, and use it to reinterpret moral disagreement, ethical theories, and moral progress within a unified computational framework.

3. Bounded Morality as Constrained Computation

This section develops a formal model of moral reasoning as a constrained computational problem. The central idea is simple: moral evaluation consists of assessing the long-run consequences of interventions in a structured system. When computational resources are limited, an agent must decide (i) how finely to represent the system and (ii) how far forward to propagate consequences. These two choices—*representational breadth* and *inferential depth*—define the core trade-off of bounded morality.

3.1. The Moral System

Moral Interaction Structure

Definition 3.1 (Moral Interaction Graph). A *moral interaction graph* is a finite graph

$$G^* = (V^*, E^*),$$

whose nodes $v \in V^*$ represent morally relevant entities and whose edges encode direct influence relationships. An edge $(u, v) \in E^*$ indicates that the morally relevant state of v may depend directly on the state or treatment of u .

Each node v has a local state space \mathcal{S}_v . The joint moral state is

$$s = (s_v)_{v \in V^*} \in \mathcal{S} := \prod_{v \in V^*} \mathcal{S}_v.$$

Let A be a finite action set. For $H \in \mathbb{N}$, define the set of action sequences

$$\mathcal{A}_H := A^H.$$

An element $\alpha = (a_0, \dots, a_{H-1}) \in \mathcal{A}_H$ represents a temporally extended intervention.

Intervention-Driven Dynamics

Definition 3.2 (Moral Dynamics). Given world state $\omega \in \Omega$ and action sequence $\alpha \in \mathcal{A}_H$, the moral state evolves as

$$s(0) = I(a_0, \omega), \quad s(t+1) = F(s(t), a_t),$$

where $F : \mathcal{S} \times A \rightarrow \mathcal{S}$ satisfies the locality condition

$$s_v(t+1) = F_v\left(s_v(t), \{s_u(t) : (u, v) \in E^*\}, a_t\right).$$

Thus interventions induce trajectories on a graph-structured dynamical system. Moral reasoning consists of evaluating these trajectories.

Ground-Truth Moral Value

Definition 3.3 (Graph-Structured Welfare). An instantaneous welfare function $U : \mathcal{S} \rightarrow \mathbb{R}$ is *graph-structured* if

$$U(s) = \sum_{v \in V^*} \psi_v(s_v) + \sum_{(u, v) \in E^*} \psi_{uv}(s_u, s_v),$$

for some node potentials $\{\psi_v\}_{v \in V^*}$ and edge potentials $\{\psi_{uv}\}_{(u, v) \in E^*}$.

Fix a discount factor $\gamma \in (0, 1)$.

Definition 3.4 (Ground-Truth Moral Value). For action sequence $\alpha \in \mathcal{A}_\infty$,

$$M^*(\alpha, \omega) := \sum_{t=0}^{\infty} \gamma^t U(s(t)).$$

The unboundedly optimal intervention is

$$\alpha^*(\omega) \in \arg \max_{\alpha} M^*(\alpha, \omega).$$

This defines the full-information moral control problem.

3.2. Bounded Representations: Breadth and Depth

In practice, agents cannot evaluate M^* exactly. They restrict both how much of the system they explicitly represent and how far forward they propagate consequences.

Abstraction and Breadth

Definition 3.5 (Abstract Moral Representation). An *abstract representation* consists of a graph

$$G = (V, E)$$

together with a surjective aggregation map

$$\pi_V : V^* \rightarrow V,$$

which partitions morally relevant entities into aggregated units.

The vertices V represent equivalence classes under π_V . Edges are induced by coarse interaction:

$$(u, v) \in E \iff \exists u' \in \pi_V^{-1}(u), \exists v' \in \pi_V^{-1}(v) \text{ s.t. } (u', v') \in E^*.$$

Thus G is the coarse-grained projection of the true moral interaction graph G^* , obtained by aggregating entities and inheriting interactions between their aggregates.

Definition 3.6 (Admissible Action Sequences Under Abstraction). Given representation G , admissible sequences are

$$\mathcal{A}_H(G) := \left\{ \alpha \in \mathcal{A}_H : a_{t,v} = \tilde{a}_{t,\pi_V(v')} \forall v' \in V^*, \forall t \right\}.$$

Thus entities aggregated together must be treated identically at every time step.

Definition 3.7 (Breadth). The *breadth* of representation G is

$$b(G) := |V| + |E|.$$

Breadth measures how many distinctions and interactions are explicitly modeled (i.e., the representational resolution).

Depth as Truncated Propagation

Definition 3.8 (Depth). The *depth* $H \in \mathbb{N}$ is a rollout horizon specifying how many steps of the dynamics are explicitly evaluated.

Definition 3.9 (Finite-Horizon Approximation). Given (G, H) , define the truncated objective

$$\hat{M}_{G,H}(\alpha, \omega) = \sum_{t=0}^H \gamma^t U(s(t)),$$

where $\alpha \in \mathcal{A}_H(G)$ and $s(t)$ evolves under the true dynamics.

The approximation error reflects only the structural constraints imposed by aggregation (G) and horizon truncation (H), not uncertainty about the underlying moral primitives or dynamics.

3.3. Informational and Inferential Costs

We associate computational cost with representational complexity and the effort required to reason over that representation.

Definition 3.10 (Total Computational Cost).

$$\text{Cost}(G, H) = \text{Cost}_{\text{info}}(G) + \text{Cost}_{\text{infer}}(G, H).$$

Definition 3.11 (Informational Cost).

$$\text{Cost}_{\text{info}}(G) = f(b(G)),$$

where f is strictly increasing in representational breadth $b(G) = |V| + |E|$.

Informational cost captures the resources required to encode the abstract interaction structure and its associated welfare and dynamical primitives at the chosen level of resolution.

Definition 3.12 (Inferential Cost).

$$\text{Cost}_{\text{infer}}(G, H) = \Theta(Hb(G)) \cdot C_{\text{search}}(G, H),$$

where $C_{\text{search}}(G, H) \geq 1$ denotes the effective number of candidate trajectories that must be evaluated to optimize action selection.

The factor $\Theta(Hb(G))$ reflects the cost of simulating a single length- H trajectory under abstraction G , while $C_{\text{search}}(G, H)$ captures the combinatorial complexity of optimizing over admissible action sequences.

3.4. Moral Strategies Under Resource Constraints

Fix a computational budget $B > 0$. Let P be a probability distribution over world states Ω .

Let \mathcal{G} denote the set of admissible abstract representations (as defined above), and let $\mathcal{A}_H(G)$ denote the admissible action sequences of length H under representation G .

Resource Allocation (Strategy Level)

Definition 3.13 (Allocation Rule). An *allocation rule* is a mapping

$$\rho : \Omega \times \mathbb{R}_+ \rightarrow \mathcal{G} \times \mathbb{N}_0,$$

such that for each (ω, B) ,

$$\rho(\omega, B) = (G, H), \quad \text{Cost}(G, H) \leq B.$$

The allocation rule determines which entities are distinguished (construction of V), which interactions are modeled (construction of E), and how far consequences are propagated (H).

Planning Within a Representation (Policy Level)

Given (G, H) and world state ω , define the induced policy

$$\pi_{G,H}(\omega) \in \arg \max_{\alpha \in \mathcal{A}_H(G)} \hat{M}_{G,H}(\alpha, \omega).$$

Bounded Moral Strategy

Definition 3.14 (Bounded Moral Strategy). A *bounded moral strategy* is specified by an allocation rule ρ . It induces a bounded moral policy

$$\pi_B^p(\omega) = \pi_{G,H}(\omega), \quad \text{where } (G, H) = \rho(\omega, B).$$

Example: Content Moderation as Bounded Morality

A minimal instantiation of the framework (full details in Appendix B).

Ground Truth. Let $G^* = (V^*, E^*)$ with $V^* = \{E_1, E_2, C, M_1, M_2\}$ (extremists E_i , connector C , moderates M_j). Polarization diffuses over G^* under sanction-modified linear dynamics with slowly decaying resentment.

Instantaneous welfare penalizes polarization magnitude and disagreement: $U(s) = -\frac{1}{5} \sum_v s_v^2 - \frac{\lambda}{6} \sum_{\{u,v\} \in E^*} (s_u - s_v)^2$, and ground-truth value is $M^*(a, \omega) = \sum_{t=0}^{\infty} \gamma^t U(s(t))$.

Competing Interventions. Sanction extremists only: $a^{(E)} = \mathbf{1}_{\{E_1, E_2\}}$, or sanction extremists and the connector: $a^{(EC)} = \mathbf{1}_{\{E_1, E_2, C\}}$.

Depth Limitation. A bounded planner evaluates $\hat{M}_{G^*, H}(a, \omega) = \sum_{t=0}^H \gamma^t U(s(t))$.

H	$a^{(E)}$	$a^{(EC)}$
2	-0.297	-0.088
10	-0.581	-0.715
∞	-0.634	-0.955

At $H = 2$, the planner selects $a^{(EC)}$. At $H = 10$, it selects $a^{(E)}$. Under the infinite-horizon objective, $a^{(E)}$ is optimal, so shallow depth incurs regret ≈ 0.322 .

Breadth Limitation. If E_1, E_2, C are aggregated into a single abstract node, admissible actions must treat them identically, eliminating $a^{(E)}$. Even with large depth, the planner cannot implement the optimal intervention, inducing the same regret magnitude.

Interpretation. Limited depth hides delayed backlash effects. Limited breadth removes selective interventions. Moral progress corresponds to reallocating computational resources across depth and breadth to reduce expected regret.

3.5. Regret and Moral Progress

Definition 3.15 (State-Dependent Regret). For world state ω , the regret of bounded moral strategy ρ under budget B is

$$R(\omega; \rho, B) = M^*(\alpha^*(\omega), \omega) - M^*(\pi_B^\rho(\omega), \omega).$$

Definition 3.16 (Expected Regret).

$$\mathbb{E}_{\omega \sim P}[R(\omega; \rho, B)].$$

A bounded moral strategy ρ is distributionally efficient under (P, B) if no other feasible strategy performs at least as well in nearly all likely world states (under P) and strictly better in some of them.

Definition 3.17 (Moral Progress). Fix a distribution P over Ω and a budget $B > 0$. Let ρ_1 and ρ_2 be feasible bounded moral strategies (i.e., $\text{Cost}(G, H) \leq B$ for all $(G, H) = \rho_i(\omega, B)$).

We say that ρ_2 exhibits *moral progress relative to* ρ_1 under (P, B) if

$$\mathbb{E}_{\omega \sim P}[R(\omega; \rho_2, B)] < \mathbb{E}_{\omega \sim P}[R(\omega; \rho_1, B)].$$

In general, progress may arise from improvements in:

1. resource allocation (selection of the abstraction (G, H)),
2. estimation (learning the welfare potentials ψ and dynamics F),
3. optimization (maximization of $\hat{M}_{G, H}$ over $\mathcal{A}_H(G)$).

In this paper, we isolate the first source and focus on progress driven by improved allocation rules ρ . We assume that the moral primitives and dynamics are known, and that optimization within a fixed representation (G, H) is exact, so that regret arises solely from structural resource allocation.

Moral Progress via Improved Resource Allocation

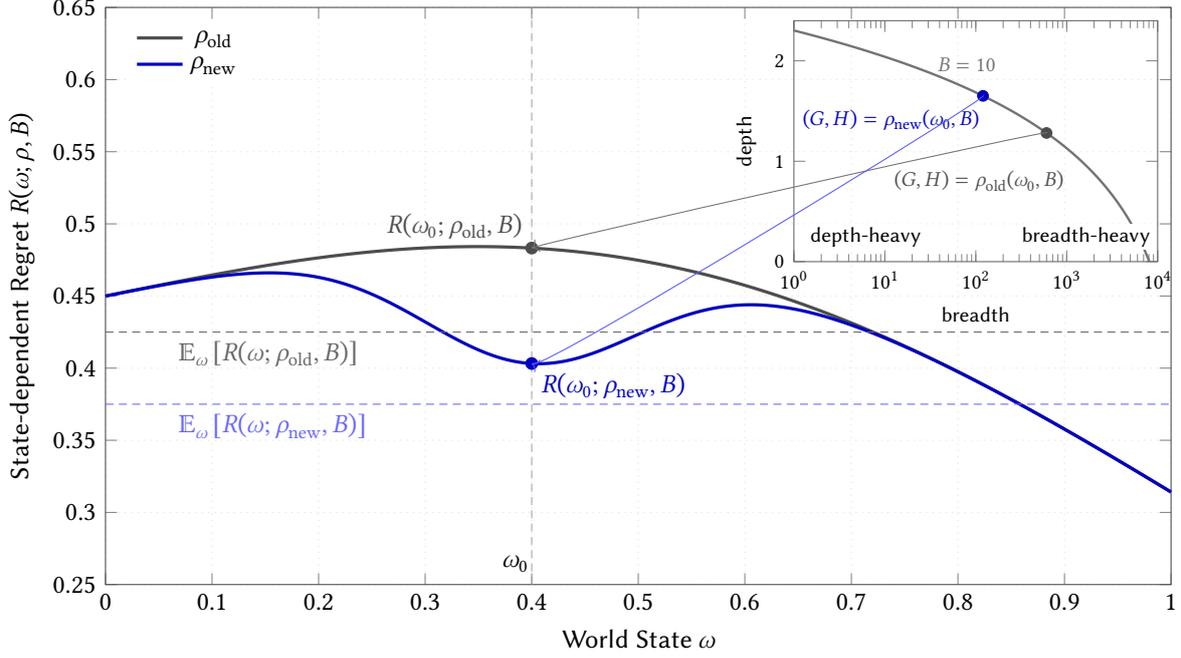


Figure 1: At fixed budget B and distribution P , policies ρ_{old} and ρ_{new} map world states to structural allocations (G, H) . The plot highlights a representative state ω_0 where the improved policy selects a different breadth–depth trade-off $(G, H) = \rho_{\text{new}}(\omega_0, B)$ on the fixed-budget frontier, yielding lower state-dependent regret $R(\omega_0; \rho_{\text{new}}, B)$ than $R(\omega_0; \rho_{\text{old}}, B)$. Dashed lines indicate expected regret $\mathbb{E}_\omega[R(\omega; \rho, B)]$ (here shown for a uniform ω).

Remark 3.18 (Relation to Causality, Planning, and Control). The formulation above places moral decision-making within the theory of controlled dynamical systems on graphs. Actions function as interventions that set initial conditions and influence local update rules, paralleling the use of do-operators in structural causal models. Moral evaluation then corresponds to planning in a dynamical system where value is assigned to entire trajectories induced by interventions.

In this interpretation, depth H plays the role of a bounded planning horizon, while abstraction (G, π_V) corresponds to state aggregation and parameter tying in approximate dynamic programming. The bounded morality problem can thus be viewed as a constrained optimal control problem in which representational resolution and rollout depth jointly restrict attainable performance.

Unlike standard control settings, however, the objective function is morally structured—it decomposes over individuals and their relationships. As a result, representational choices are ethically consequential. Coarse aggregation or truncated propagation can induce moral error even when the underlying welfare function is correct.

3.6. Asymptotic Scaling and a Canonical Cost Model

We now conjecture asymptotic bounds on informational and inferential costs, introduce a tractable canonical model consistent with these bounds, and derive the resulting breadth–depth trade-off under a fixed computational budget.

Asymptotic Bounds

Let $b = b(G) = |V| + |E|$.

Proposition 3.19 (Informational Lower Bound). *There exists $c_1 > 0$ such that*

$$\text{Cost}_{\text{info}}(G) \geq c_1 b.$$

Proof. Any abstraction distinguishing $|V|$ aggregates and $|E|$ interactions must encode at least these structural elements. \square

Proposition 3.20 (Inferential Lower Bound). *There exists $c_2 > 0$ such that*

$$\text{Cost}_{\text{infer}}(G, H) \geq c_2 Hb.$$

Proof. Simulating a single trajectory of length H requires $\Theta(b)$ work per step under locality, yielding $\Theta(Hb)$ total effort. \square

Evaluating a single trajectory scales linearly in breadth and depth. Planning, however, may require comparing exponentially many trajectories in the worst case. When interactions are sufficiently localized, this growth can instead remain polynomial in the horizon. Thus

$$\Omega(Hb) \leq \text{Cost}_{\text{infer}}(G, H) \leq O(Hb C_{\text{search}}(G, H)),$$

where $C_{\text{search}}(G, H) \geq 1$ captures search complexity.

Canonical Cost Model

To isolate the structural trade-off, we adopt a tractable model that respects the lower bounds and abstracts from worst-case search effects.

Definition 3.21 (Canonical Cost Model). For $b = b(G)$,

$$\text{Cost}(b, H) = \alpha b + \beta b H^p,$$

where $\alpha, \beta > 0$ and $p \geq 1$.

The first term models representational cost; the second models rollout effort growing linearly in breadth and polynomially in depth.

Budget Constraint and Feasible Region

Fix budget $B > 0$. Admissible pairs satisfy

$$\alpha b + \beta b H^p \leq B.$$

For $b > 0$,

$$H^p \leq \frac{B}{\beta b} - \frac{\alpha}{\beta}.$$

Whenever $B > \alpha b$, this yields

$$H \leq \left(\frac{B}{\beta b} - \frac{\alpha}{\beta} \right)^{1/p}.$$

For $B \gg \alpha b$,

$$H_{\text{max}}(b) = \Theta(b^{-1/p}).$$

Breadth–Depth Trade-Off

Corollary 3.22 (Inverse Scaling Law). *Under the canonical cost model with fixed B and $p \geq 1$, there exists $C > 0$ such that*

$$H_{\text{max}}(b) \leq C b^{-1/p}.$$

Proof. From $\alpha b + \beta b H^p \leq B$, we obtain $H^p \leq \frac{B}{\beta b} - \frac{\alpha}{\beta}$. For b sufficiently small relative to B/α , the right-hand side is bounded above by $C' b^{-1}$ for some $C' > 0$. Taking p -th roots gives $H \leq C b^{-1/p}$. \square

This corollary formalizes the structural tension at the heart of bounded morality: increasing representational resolution strictly reduces the horizon over which consequences can be propagated under fixed computational resources. The feasible region in (b, H) space is therefore downward-sloping, and any allocation of finite resources must lie on or below this breadth–depth frontier.

4. Ethical Theories as Resource-Bounded Moral Strategies

“We do not grow absolutely, chronologically. We grow sometimes in one dimension, and not in another; unevenly. We are relative. We are mature in one realm, childish in another.”

– Anaïs Nin

Section 3 formalized moral reasoning as a constrained computational problem: under a budget B , an allocation rule ρ selects an abstract representation (G, π_V) and a rollout depth H , after which the agent optimizes a truncated objective over admissible intervention sequences $\alpha \in \mathcal{A}_H(G)$. Regret arises when structural restrictions on breadth and depth prevent recovery of the ground-truth optimum.

In this section, we reinterpret several canonical ethical theories as *families of bounded moral strategies*. Each theory is modeled as a structured restriction on:

1. admissible allocation rules ρ (hence on feasible (G, H) pairs),
2. the aggregation functional used to evaluate trajectories,
3. and, where relevant, additional admissibility constraints on interventions.

The underlying moral primitives (G^*, F, U) remain fixed. Differences between theories are represented as differences in how computational resources are allocated and how trajectory evaluations are constructed under constraint.

The mappings below are deliberately schematic. They isolate dominant structural commitments—breadth, depth, aggregation, and admissibility—rather than attempting to reconstruct full philosophical doctrines.

4.1. Strategy Families

Let $(G, H) = \rho(\omega, B)$ and let $s(0:H)$ denote the trajectory induced by $\alpha \in \mathcal{A}_H(G)$ under the true dynamics. A strategy family is characterized by an evaluation functional

$$\mathcal{T}\left(\{s(t)\}_{t=0}^H; G\right),$$

and (optionally) a restricted admissible class $\mathcal{A}_H^{\text{adm}}(G) \subseteq \mathcal{A}_H(G)$. The induced policy is

$$\pi_B^{\rho, \mathcal{T}}(\omega) \in \arg \max_{\alpha \in \mathcal{A}_H^{\text{adm}}(G)} \mathcal{T}\left(\{s(t)\}_{t=0}^H; G\right).$$

Under the baseline model of Section 3, \mathcal{T} coincides with the truncated discounted sum $\hat{M}_{G,H}$ and $\mathcal{A}_H^{\text{adm}}(G) = \mathcal{A}_H(G)$. Ethical theories can be represented as structured deviations from this baseline.

4.2. Act Utilitarianism

Act utilitarianism evaluates interventions by maximizing total welfare across all affected individuals [43]. In the present formalism, its defining commitment is additive aggregation over a broad representational scope.

Aggregation. The evaluation functional coincides with the truncated ground-truth objective:

$$\mathcal{T}_{\text{UTIL}}\left(\{s(t)\}_{t=0}^H; G\right) = \sum_{t=0}^H \gamma^t U(s(t)).$$

Theory	Breadth	Depth	Structural Commitment	Typical Regime
Act Utilitarianism	High	Mod.	Additive sum $\sum_t \gamma^t U(s(t))$	Large, weakly coupled systems
Deontology	Mod.	Mod.	Hard constraints on $\mathcal{A}_H(G)$	Tight rules; limited scope
Contractualism	Low	High	Worst-case $\min_v \sum_t \gamma^t U_v(s(t))$	Few parties; deep objections
Virtue Ethics	Very Low	0	Amortized policy π_θ	Real-time action
Care Ethics	Local	Mod.	Relational (edge-weighted) terms	Dense local dependence

Table 2

Ethical theories as caricatured families of bounded moral strategies. Each theory occupies a characteristic region of the breadth–depth frontier and imposes a distinct structural commitment on evaluation or admissibility.

Allocation structure. A utilitarian strategy family consists of allocation rules ρ that prioritize representational breadth:

$$\mathcal{R}_{\text{UTIL}} = \left\{ \rho : \rho(\omega, B) = (G, H) \text{ with } b(G) \text{ maximized under } B, H \text{ moderate} \right\}.$$

Depth is limited by budget but not minimized; breadth is expanded to include as many morally relevant entities and interactions as feasible.

Computational interpretation. Computation is devoted primarily to representing many stakeholders. Aggregation remains structurally simple. This strategy is locally efficient when interaction effects are weak or approximately decomposable, but becomes expensive when deep dependencies dominate.

4.3. Deontological Rule-Based Ethics

Deontological theories evaluate interventions by compliance with rules or duties rather than by outcome maximization [44]. In the present framework, this corresponds primarily to restricting admissible interventions.

Admissibility and search reduction. Let \mathcal{C} denote a set of rule constraints. Define

$$\mathcal{A}_{H, \mathcal{C}}^{\text{adm}}(G) = \{ \alpha \in \mathcal{A}_H(G) : \alpha \text{ satisfies } \mathcal{C} \}.$$

The evaluation functional imposes a hard barrier:

$$\mathcal{J}_{\text{RULE}}(\{s(t)\}; G) = \begin{cases} \sum_{t=0}^H \gamma^t U(s(t)), & \alpha \in \mathcal{A}_{H, \mathcal{C}}^{\text{adm}}(G), \\ -\infty, & \text{otherwise.} \end{cases}$$

The constraints reduce the effective search space from $\mathcal{A}_H(G)$ to the feasible subset $\mathcal{A}_{H, \mathcal{C}}^{\text{adm}}(G)$, potentially yielding large savings in the search factor $C_{\text{search}}(G, H)$.

However, feasibility typically decreases as breadth and depth increase. Larger $b(G)$ introduces more entities and constraint instances; larger H requires constraints to hold over longer trajectories. In general,

$$\mathcal{A}_{H, \mathcal{C}}^{\text{adm}}(G) = \bigcap_{t=0}^H \{ \alpha : \text{constraints hold at time } t \},$$

so the admissible set shrinks with either $b(G)$ or H . In extreme cases it may be empty, corresponding to moral infeasibility under the chosen scope and horizon.

Allocation structure. A canonical deontological family therefore restricts both breadth and depth:

$$\mathcal{R}_{\text{RULE}} = \left\{ \rho : \rho(\omega, B) = (G, H) \text{ with } b(G) \text{ and } H \text{ medium-to-small} \right\}.$$

Computational interpretation. Deontological strategies gain efficiency by pruning the search space rather than simplifying aggregation. Their tractability depends on maintaining medium breadth and depth; as scope or horizon expands, constraint-checking costs grow and the feasible set can collapse.

4.4. Contractualism

Contractualist theories evaluate interventions by whether they can be justified to each affected individual, often modeled via the strongest reasonable objection [45]. The defining structural feature is worst-case aggregation over represented individuals.

Aggregation. Let $U_v(s)$ denote the portion of instantaneous welfare attributed to represented entity $v \in V$ under abstraction G . Define

$$\mathcal{T}_{\text{CONTRACT}}(\{s(t)\}_{t=0}^H; G) = \min_{v \in V} \sum_{t=0}^H \gamma^t U_v(s(t)).$$

Allocation structure. Contractualist strategy families restrict breadth while increasing depth:

$$\mathcal{R}_{\text{CONTRACT}} = \left\{ \rho : \rho(\omega, B) = (G, H) \text{ with } b(G) \text{ small, } H \text{ large under } B \right\}.$$

Computational interpretation. Attention is concentrated on a limited set of stakeholders, but substantial depth is allocated to propagate indirect consequences and objections. This strategy is viable in small-scale, high-stakes settings with sufficient deliberative resources.

4.5. Virtue Ethics

Virtue ethics emphasizes stable dispositions rather than explicit outcome calculation [46]. In the present model, this corresponds to amortized inference.

Policy form. Set $H = 0$ and define a learned policy

$$\pi_\theta : \Omega \rightarrow A,$$

where parameters θ encode dispositions acquired through experience.

Allocation structure. The associated strategy family satisfies

$$\mathcal{R}_{\text{VIRTUE}} = \left\{ \rho : \rho(\omega, B) = (G, 0) \text{ with } b(G) \text{ minimal} \right\}, \quad \pi_B^\rho(\omega) = \pi_\theta(\omega).$$

Computational interpretation. Online rollout is eliminated. Computation is shifted offline into learning. This strategy is efficient under extreme time constraints but sensitive to context.

4.6. Care Ethics

Care ethics emphasizes concrete relationships and contextual interdependence [47]. Its defining feature in the present framework is localized representation of dense relational structure.

Representation. Let v_0 be a focal agent and let $\mathcal{N}_r(v_0)$ denote a radius- r neighborhood in G^* . Define

$$G \approx G^*[\mathcal{N}_r(v_0)].$$

Ethical Theories as Moral Strategy Families in Breadth–Depth Space

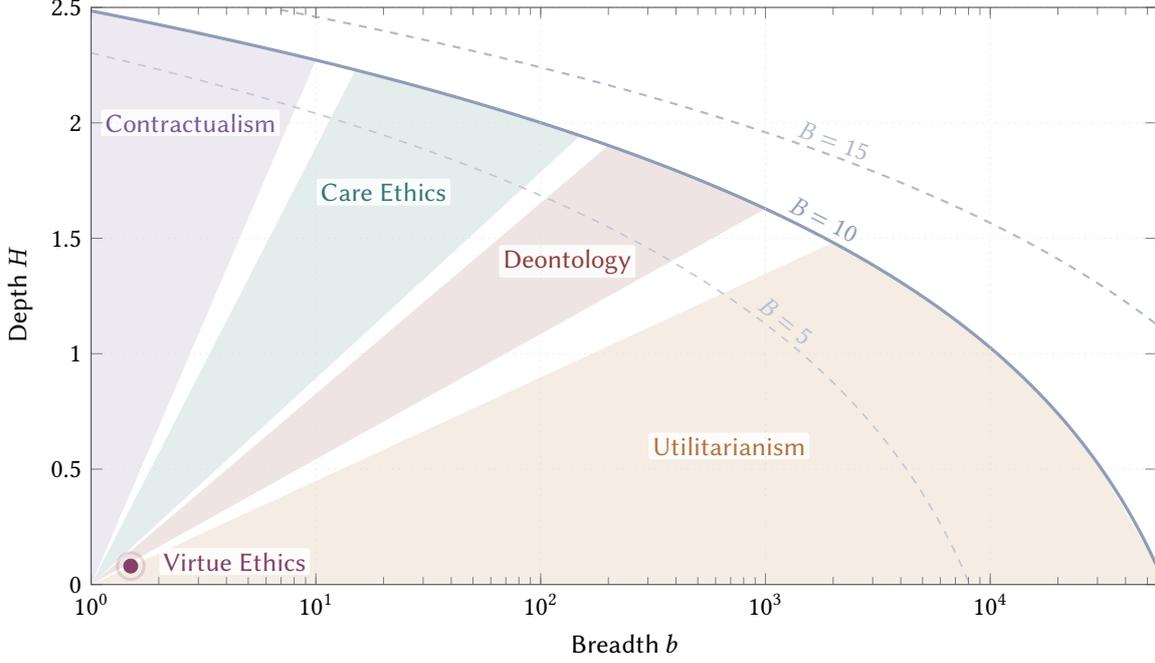


Figure 2: Budget contours (B) induce Pareto-efficient breadth–depth frontiers under a canonical cost model. Ethical theories are depicted as strategy families with characteristic scaling tendencies: utilitarianism trades budget for breadth, contractualism for depth, care ethics for local breadth and moderate depth, while deontology and virtue ethics remain near low-depth regimes. Regions are conceptual caricatures.

Aggregation. Evaluation emphasizes relational (edge) terms:

$$\mathcal{T}_{\text{CARE}}(\{s(t)\}_{t=0}^H; G) = \sum_{t=0}^H \gamma^t \sum_{(u,v) \in E} \psi_{uv}(s_u(t), s_v(t)).$$

Allocation structure.

$$\mathcal{R}_{\text{CARE}} = \left\{ \rho : \rho(\omega, B) = (G, H) \text{ with } G \text{ local around } v_0, H \text{ moderate} \right\}.$$

Computational interpretation. Representational capacity is devoted to dense local interactions rather than global enumeration. Depth is sufficient to integrate relational dependencies within the local subgraph but does not scale globally.

4.7. Ethical Disagreement as Allocation Disagreement

On this formalization, ethical disagreement can arise from both disagreement about welfare primitives but also from disagreement about resource allocation. Distinct theories correspond to different structured restrictions on (G, H) , different aggregation functionals \mathcal{T} , and different admissibility constraints. Under fixed budget B and distribution P , these structural commitments induce different regret profiles.

Ethical theories can thus be interpreted as locally efficient regions of the breadth–depth frontier (Figure 2). Each represents a characteristic solution to the problem of allocating limited computational resources across representational scope and inferential depth in morally structured dynamical systems.

5. Discussion

This paper advances a limited but we believe, clarifying perspective on moral reasoning: even if moral truth is well defined, reasoning about it is a constrained computational problem. By making

representational scope, inferential depth, and computational costs explicit, the framework highlights structural tradeoffs that any finite moral agent—human or artificial—must confront.

A first implication is a reframing of moral failure. Many familiar shortcomings in moral judgment—oversimplification, rigidity, or neglect of distant stakeholders—need not be attributed to defective values or irrationality. Instead, they can arise from rational allocations of limited resources across representation and inference. The breadth–depth tradeoff implies that attending to more entities necessarily constrains how deeply their interactions can be reasoned about. This does not excuse moral failings, but it shifts explanatory weight toward underlying computational constraints rather than attributing outcomes solely to character or principle.

Second, the framework offers a non-relativist account of persistent moral disagreement. Agents who share the same underlying moral objective may nonetheless reach different conclusions because they operate at different feasible points in the breadth–depth space or face different resource constraints. Disagreement, on this view, can reflect differences in approximation rather than differences in moral values. Persistent moral disagreement therefore need not imply moral pluralism, but may instead reflect stable variation in how agents approximate a common moral objective under constraint.

Third, the results offer a reinterpretation of ethical theories. Rather than treating utilitarianism, contractualism, or care ethics as competing foundational accounts, the framework treats them as families of reasoning strategies that are locally efficient under different constraints. Broad but shallow aggregation, narrow but deep reasoning, and heuristic rule-following each occupy different regions of the feasible space. This perspective does not adjudicate between theories, but it helps explain why different approaches appear compelling in different contexts.

The framework also introduces a new notion of moral progress. Progress need not consist in expanding the scope of moral concern, discovering new values, or converging on a single moral doctrine. Instead, progress can occur through improved abstractions, more efficient inference procedures, better data, or better allocation of moral reasoning effort across contexts—each of which reduces regret under fixed resources. This reframes historical and institutional moral change as, in part, advances in how moral reasoning is computed.

Finally, the implications for artificial systems are primarily diagnostic rather than prescriptive. Moral competence should not be reduced to a single objective score, but instead evaluated relative to resource budgets and deployment contexts along dimensions of scope, depth, and latency. Many alignment failures may reflect mismatches between moral computation and context rather than incorrect objectives. The framework provides a technical vocabulary for analyzing where and why morally competent behavior breaks down under constraint.

Overall, Bounded Morality does not aim to compete with moral philosophy or moral psychology. Its contribution is narrower: to isolate a structural feature that any account of moral reasoning must contend with when instantiated in finite agents, and to show how this feature shapes disagreement, heuristics, and the limits of moral deliberation.

6. Conclusion

This paper introduced *Bounded Morality*, a framework that treats moral reasoning as constrained inference over structured moral worlds. By explicitly modeling representational scope, inferential depth, and computational cost, we showed that finite agents face an unavoidable tradeoff between moral breadth and moral depth. This tradeoff constrains which moral strategies are achievable under fixed resources, making explicit the limits of moral reasoning for finite agents. Conceptually, the framework reframes moral disagreement, heuristics, and ethical theories as predictable consequences of bounded computation rather than as failures of moral concern or coherence. For artificial systems, it highlights that alignment is not only a matter of specifying correct values, but also of designing moral computation that is well matched to capacity and context. More broadly, Bounded Morality provides a minimal computational lens through which insights from ethics, psychology, and AI can be integrated and evaluated under realistic constraints.

Acknowledgments

This work was supported by the Amaranth Foundation. The authors thank Jared Moore and David Gottlieb for valuable discussions, including insights drawn from their instruction in Stanford's CS186: How to Make a Moral Agent course, and Nicholas Christakis for helpful conversations.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 in order to: Improve writing style, Content enhancement. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] M. Takeshita, R. Rafal, K. Araki, Towards Theory-based Moral AI: Moral AI with Aggregating Models Based on Normative Ethical Theory, 2023. URL: <http://arxiv.org/abs/2306.11432>. doi:10.48550/arXiv.2306.11432, arXiv:2306.11432 [cs].
- [2] A. Hegde, V. Agarwal, S. Rao, Ethics, Prosperity, and Society: Moral Evaluation Using Virtue Ethics and Utilitarianism, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 2020, pp. 167–174. URL: <https://www.ijcai.org/proceedings/2020/24>. doi:10.24963/ijcai.2020/24.
- [3] G. R. T. White, A. Samuel, P. Jones, N. Madhavan, A. Afolayan, A. Abdullah, T. Kaushik, Mapping the ethic-theoretical foundations of artificial intelligence research, *Thunderbird International Business Review* 66 (2024) 171–183. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/tie.22368>. doi:10.1002/tie.22368, eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/tie.22368>.
- [4] V. Preniqi, I. Ghinassi, J. Ive, C. Saitis, K. Kalimeri, MoralBERT: A Fine-Tuned Language Model for Capturing Moral Values in Social Discussions, in: Proceedings of the 2024 International Conference on Information Technology for Social Good, ACM, Bremen Germany, 2024, pp. 433–442. URL: <https://dl.acm.org/doi/10.1145/3677525.3678694>. doi:10.1145/3677525.3678694.
- [5] I. Gabriel, Artificial intelligence, values, and alignment, *Minds and machines* 30 (2020) 411–437.
- [6] H. A. Simon, Bounded Rationality, in: J. Eatwell, M. Milgate, P. Newman (Eds.), *Utility and Probability*, Palgrave Macmillan UK, London, 1990, pp. 15–18. URL: https://doi.org/10.1007/978-1-349-20568-4_5. doi:10.1007/978-1-349-20568-4_5.
- [7] J. R. Anderson, *The Adaptive Character of Thought*, 1 ed., Psychology Press, 2013. URL: <https://www.taylorfrancis.com/books/9780203771730>. doi:10.4324/9780203771730.
- [8] T. L. Griffiths, F. Lieder, N. D. Goodman, Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic, *Topics in Cognitive Science* 7 (2015) 217–229. URL: <https://onlinelibrary.wiley.com/doi/10.1111/tops.12142>. doi:10.1111/tops.12142.
- [9] S. Levine, N. Chater, J. B. Tenenbaum, F. Cushman, Resource-rational contractualism: A triple theory of moral cognition, *Behavioral and Brain Sciences* (2024) 1–38. URL: <https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/resourcerational-contractualism-a-triple-theory-of-moral-cognition/5A567D41A472DBC0D965460966580C74>. doi:10.1017/S0140525X24001067.
- [10] S. Levine, M. Franklin, T. Zhi-Xuan, S. Y. Guyot, L. Wong, D. Kilov, Y. Choi, J. B. Tenenbaum, N. Goodman, S. Lazar, Resource Rational Contractualism Should Guide AI Alignment, *arXiv preprint arXiv:2506.17434* (2025).
- [11] D. Chugh, M. H. Bazerman, M. R. Banaji, Bounded Ethicality as a Psychological Barrier to Recognizing Conflicts of Interest, in: *Conflicts of interest: Challenges and solutions in business, law, medicine, and public policy*, Cambridge University Press, New York, NY, US, 2005, pp. 74–95. doi:10.1017/CBO9780511610332.006.

- [12] A. E. Tenbrunsel, K. Smith-Crowe, 13 Ethical Decision Making: Where We've Been and Where We're Going, *Academy of Management Annals* 2 (2008) 545–607. URL: <http://journals.aom.org/doi/10.5465/19416520802211677>. doi:10.5465/19416520802211677.
- [13] D. Marr, *Vision: a computational investigation into the human representation and processing of visual information*, MIT Press, Cambridge, Mass, 2010.
- [14] J. Piaget, M. Cook, *The origins of intelligence in children*, volume 8, International universities press New York, 1952. Issue: 5.
- [15] L. Kohlberg, *Moral stages and moralization: The cognitive-development approach*, *Moral development and behavior: Theory research and social issues* (1976) 31–53. Publisher: Rinehart & Winston.
- [16] N. Eisenberg, R. A. Fabes, T. L. Spinrad, *Prosocial Development*, in: *Handbook of child psychology: Social, emotional, and personality development*, Vol. 3, 6th ed, John Wiley & Sons, Inc., Hoboken, NJ, US, 2006, pp. 646–718.
- [17] C. R. Crimston, P. G. Bain, M. J. Hornsey, B. Bastian, *Moral expansiveness: Examining variability in the extension of the moral world.*, *Journal of personality and social psychology* 111 (2016) 636. Publisher: American Psychological Association.
- [18] P. Singer, *The expanding circle*, Clarendon Press Oxford, 1981.
- [19] D. Parfit, *Reasons and Persons*, 1 ed., Oxford University Press Oxford, 1986. URL: <https://academic.oup.com/book/12484>. doi:10.1093/019824908X.001.0001.
- [20] S. M. Gardiner, *A perfect moral storm: The ethical tragedy of climate change*, Oxford University Press, 2011.
- [21] A. Waytz, J. Cacioppo, N. Epley, *Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism*, *Perspectives on Psychological Science* 5 (2010) 219–232. URL: <https://journals.sagepub.com/doi/10.1177/1745691610369336>. doi:10.1177/1745691610369336.
- [22] J. Haidt, *The emotional dog and its rational tail: A social intuitionist approach to moral judgment*, *Psychological Review* 108 (2001) 814–834. doi:10.1037/0033-295X.108.4.814, place: US Publisher: American Psychological Association.
- [23] J. D. Greene, L. E. Nystrom, A. D. Engell, J. M. Darley, J. D. Cohen, *The Neural Bases of Cognitive Conflict and Control in Moral Judgment*, *Neuron* 44 (2004) 389–400. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0896627304006348>. doi:10.1016/j.neuron.2004.09.027.
- [24] J. D. Greene, S. A. Morelli, K. Lowenberg, L. E. Nystrom, J. D. Cohen, *Cognitive load selectively interferes with utilitarian moral judgment*, *Cognition* 107 (2008) 1144–1154. doi:10.1016/j.cognition.2007.11.004.
- [25] P. Conway, B. Gawronski, *Deontological and utilitarian inclinations in moral decision making: a process dissociation approach*, *Journal of Personality and Social Psychology* 104 (2013) 216–235. doi:10.1037/a0031021.
- [26] R. L. Selman, *The Growth of Interpersonal Understanding: Developmental and Clinical Analyses*, Academic Press, 1980.
- [27] J. Piaget, *The moral judgment of the child*, Routledge, 2013.
- [28] M. Buon, P. Jacob, E. Loissel, E. Dupoux, *A non-mentalistic cause-based heuristic in human social evaluations*, *Cognition* 126 (2013) 149–155. doi:10.1016/j.cognition.2012.09.006.
- [29] J. W. Martin, M. Buon, F. Cushman, *The Effect of Cognitive Load on Intent-Based Moral Judgment*, *Cognitive Science* 45 (2021) e12965. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12965>. doi:10.1111/cogs.12965, _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12965>.
- [30] J. D. Greene, *Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics*, *Ethics* 124 (2014) 695–726. URL: <https://www.journals.uchicago.edu/doi/10.1086/675875>. doi:10.1086/675875.
- [31] G. Kahane, J. A. C. Everett, B. D. Earp, L. Caviola, N. S. Faber, M. J. Crockett, J. Savulescu, *Beyond sacrificial harm: A two-dimensional model of utilitarian psychology*, *Psychological Review* 125 (2018) 131–164. doi:10.1037/rev0000093.
- [32] D. FETHERSTONHAUGH, P. SLOVIC, S. JOHNSON, J. FRIEDRICH, *Insensitivity to the Value*

- of Human Life: A Study of Psychophysical Numbing, *Journal of Risk and Uncertainty* 14 (1997) 283–300. URL: <https://doi.org/10.1023/A:1007744326393>. doi:10.1023/A:1007744326393.
- [33] R. S. Suter, R. Hertwig, Time and moral judgment, *Cognition* 119 (2011) 454–458. doi:10.1016/j.cognition.2011.01.018.
- [34] J. M. Paxton, L. Ungar, J. D. Greene, Reflection and reasoning in moral judgment, *Cognitive Science* 36 (2012) 163–177. doi:10.1111/j.1551-6709.2011.01210.x.
- [35] J. Baron, M. Spranca, Protected Values, *Virology* 70 (1997) 1–16. doi:10.1006/obhd.1997.2690.
- [36] S. Dickert, D. Västfjäll, J. Kleber, P. Slovic, Scope insensitivity: The limits of intuitive valuation of human lives in public policy, *Journal of Applied Research in Memory and Cognition* 4 (2015) 248–255. URL: <https://www.sciencedirect.com/science/article/pii/S2211368114000795>. doi:10.1016/j.jarmac.2014.09.002.
- [37] F. Brandt (Ed.), *Handbook of computational social choice*, Cambridge University Press, Cambridge ; New York, 2016.
- [38] K. Arrow, *Social Choice and Individual Values*, Cowles Foundation Monograph Series, Yale University Press, 1970. URL: <https://books.google.com/books?id=uebtAAAAMAAJ>.
- [39] V. Conitzer, T. Sandholm, Communication complexity of common voting rules, in: *Proceedings of the 6th ACM conference on Electronic commerce*, ACM, Vancouver BC Canada, 2005, pp. 78–87. URL: <https://dl.acm.org/doi/10.1145/1064009.1064018>. doi:10.1145/1064009.1064018.
- [40] V. Conitzer, R. Freedman, J. Heitzig, W. H. Holliday, B. M. Jacobs, N. Lambert, M. Mossé, E. Pacuit, S. Russell, H. Schoelkopf, E. Tewolde, W. S. Zwicker, *Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback*, 2024. URL: <http://arxiv.org/abs/2404.10271>. doi:10.48550/arXiv.2404.10271, arXiv:2404.10271 [cs].
- [41] G. E. Gigerenzer, R. E. Hertwig, T. E. Pachur, *Heuristics: The foundations of adaptive behavior*, Oxford university press, 2011.
- [42] C. R. Sunstein, Social Norms and Social Roles, *Columbia Law Review* 96 (1996) 903. URL: <https://www.jstor.org/stable/1123430?origin=crossref>. doi:10.2307/1123430.
- [43] J. S. Mill, *Utilitarianism*, in: *Seven masterpieces of philosophy*, Routledge, 2016, pp. 329–375.
- [44] I. Kant, *Groundwork of the metaphysic of morals*, in: *Immanuel Kant*, Routledge, 2020, pp. 17–98.
- [45] T. M. Scanlon, *What we owe to each other*, Belknap Press, 2000.
- [46] P. Gottlieb, *Aristotle: nicomachean ethics*, in: *Central Works of Philosophy v1*, Routledge, 2015, pp. 46–68.
- [47] C. Gilligan, *In a different voice: Psychological theory and women’s development*, Harvard university press, 1993.

A. Formal Definitions and Notation

Symbol / Term	Name	Definition / Role
$G^* = (V^*, E^*)$	Moral interaction graph	Full graph of morally relevant entities and their local influence relations.
$v \in V^*$	Moral entity	Node whose state contributes directly to moral welfare.
\mathcal{S}_v	Local state space	State space associated with entity v .
A	Action set	Finite set of atomic interventions available at each time step.
$\mathcal{A}_H = A^H$	Action sequences	Length- H intervention sequences (temporally extended control policies).
F	Moral dynamics	Local graph-structured state update rule governing trajectory evolution.
$U(s)$	Instantaneous welfare	Graph-structured welfare function decomposing into node and edge potentials.
$M^*(\alpha, \omega)$	Ground-truth moral value	Infinite-horizon discounted welfare induced by action sequence α .
$\pi_V : V^* \rightarrow V$	Aggregation map	Surjection defining coarse-grained moral representation.
$G = (V, E)$	Abstract moral representation	Coarse interaction graph induced by aggregation.
$b(G) = V + E $	Breadth	Representational complexity of abstraction G .
H	Depth	Rollout horizon for consequence propagation.
$\hat{M}_{G,H}$	Truncated objective	Finite-horizon approximation to M^* under abstraction G .
$\text{Cost}_{\text{info}}(G)$	Informational cost	Cost of encoding representation G ; increasing in $b(G)$.
$\text{Cost}_{\text{infer}}(G, H)$	Inferential cost	Cost of simulating and optimizing over length- H trajectories under G .
$\text{Cost}(G, H)$	Total cost	Sum of informational and inferential costs.
B	Resource budget	Upper bound on allowable computational cost.
ρ	Allocation rule	Strategy mapping (ω, B) to feasible (G, H) .
π_B^ρ	Bounded moral policy	Intervention sequence selected under allocation rule ρ .
$R(\omega; \rho, B)$	State-dependent regret	Loss in ground-truth value relative to optimal unbounded intervention.
P	World distribution	Distribution over world states used to evaluate expected regret.

Table 3

Comprehensive glossary of formal objects used in the Bounded Morality framework. This table consolidates the mathematical definitions introduced in Section 3 for reference.

B. A Worked Example of Bounded Morality

We illustrate the Bounded Morality framework with a minimal model of content moderation that instantiates the moral interaction graph, intervention-driven dynamics, graph-structured welfare, abstraction, and regret defined in Section 3. The purpose is to show explicitly how a concrete policy problem maps onto the formal objects of the theory. The example demonstrates that computational constraints—specifically limited inferential depth and limited representational breadth—can invert moral decisions even when the welfare function and dynamics are fully known and optimization within a fixed representation is exact.

Two phenomena emerge:

1. **Depth Reversal:** Truncated evaluation (H small) favors aggressive intervention, while longer rollouts (H large) favor targeted intervention.
2. **Breadth Constraint:** Coarse abstraction can eliminate the targeted intervention from the admissible action set.

B.1. Instantiation of the Moral System

Moral Interaction Graph. Let

$$V^* = \{E_1, E_2, C, M_1, M_2\},$$

representing two extremist nodes, a connector node, and two moderate nodes.

The moral interaction graph $G^* = (V^*, E^*)$ is

$$E^* = \{\{E_1, E_2\}, \{E_1, C\}, \{E_2, C\}, \{C, M_1\}, \{C, M_2\}, \{M_1, M_2\}\}.$$

The connector C is the unique pathway through which polarization flows from extremists to moderates.

State Space. Each node $v \in V^*$ has polarization state $s_v(t) \in \mathbb{R}$ and resentment state $r_v(t) \in \mathbb{R}$. The full moral state is $s(t) = (s_v(t))_{v \in V^*} \in \mathbb{R}^5$.

Actions. An intervention is a sanction vector $a \in A = \{0, 1\}^5$, applied at $t = 0$ only. Thus, although the general framework allows $\alpha \in A^H$, here we reduce $\mathcal{A}_H(G^*)$ to a single sanction decision.

Sanctioning has two effects:

$$s_v(0) = \begin{cases} s_{\text{san}} & a_v = 1, \\ \omega_v & a_v = 0, \end{cases} \quad (1)$$

$$r(0) = r_0 a. \quad (2)$$

Sanctioning reduces polarization immediately but induces resentment.

Intervention-Driven Dynamics. Let W be the adjacency matrix of G^* with unit edge weights and self-loops. Define a sanction-modulated influence matrix:

$$W(a) = W \text{diag}(w(a)), \quad w_u(a) = \begin{cases} k & a_u = 1, \\ 1 & a_u = 0, \end{cases}$$

and normalize rows to obtain $P(a)$.

Polarization and resentment evolve as:

$$s(t+1) = \beta P(a)s(t) + \alpha r(t), \quad (3)$$

$$r(t+1) = (1 - \delta)r(t). \quad (4)$$

Because $P(a)$ is row-stochastic and $\beta < 1$, the spectral radius of $\beta P(a)$ is strictly less than one, i.e., the system is globally stable. Reversal effects therefore arise from truncation rather than instability.

Graph-Structured Welfare. Let $n = |V^*|$ and $m = |E^*|$. Define instantaneous welfare:

$$U(s) = -\frac{1}{n} \sum_{v \in V^*} s_v^2 - \frac{\lambda}{m} \sum_{\{u,v\} \in E^*} (s_u - s_v)^2,$$

The first term penalizes polarization magnitude; the second penalizes disagreement across edges. Ground-truth moral value is

$$M^*(a, \omega) = \sum_{t=0}^{\infty} \gamma^t U(s(t)).$$

Finite-Horizon Approximation. A bounded planner with depth H evaluates:

$$\hat{M}_{G^*,H}(a, \omega) = \sum_{t=0}^H \gamma^t U(s(t)).$$

Truncation introduces approximation error relative to M^* .

B.2. Numerical Illustration

Parameters:

$$\begin{aligned} \beta &= 0.9, & \alpha &= 0.2, \\ \delta &= 0.05, & r_0 &= 1.0, \\ k &= 0.1, & \gamma &= 0.9, \\ \lambda &= 0.1, & s_{\text{san}} &= 0.2. \end{aligned}$$

Initial polarization:

$$\omega_{E_1} = \omega_{E_2} = 1, \quad \omega_C = 0.6, \quad \omega_{M_1} = \omega_{M_2} = 0.$$

Candidate actions:

$$a^{(0)} = \mathbf{0}, \quad a^{(E)} = \mathbf{1}_{\{E_1, E_2\}}, \quad a^{(EC)} = \mathbf{1}_{\{E_1, E_2, C\}}.$$

H	$a^{(0)}$	$a^{(E)}$	$a^{(EC)}$
2	-0.765	-0.297	-0.088
10	-1.304	-0.581	-0.715
∞	-1.348	-0.634	-0.955

Table 4

Finite-horizon value $\hat{M}_{G^*,H}(a, \omega)$ and infinite-horizon value $M^*(a, \omega)$ (approximated using rollout to 2000 steps). Bold indicates the maximizer.

- At depth $H = 2$, sanctioning extremists and the connector is optimal: $\pi_{G^*,2}(\omega) = a^{(EC)}$.
- At depth $H = 10$, sanctioning extremists only is optimal: $\pi_{G^*,10}(\omega) = a^{(E)}$.
- Under the infinite-horizon objective M^* , sanctioning extremists only is optimal: $a^*(\omega) = a^{(E)}$.

The depth-2 planner instead selects $a^{(EC)}$, yielding state-dependent regret

$$R(\omega; \rho^{(2)}, B) = M^*(a^{(E)}, \omega) - M^*(a^{(EC)}, \omega) \approx 0.322.$$

By contrast, the depth-10 planner selects $a^{(E)}$ and therefore incurs zero regret within this action class.

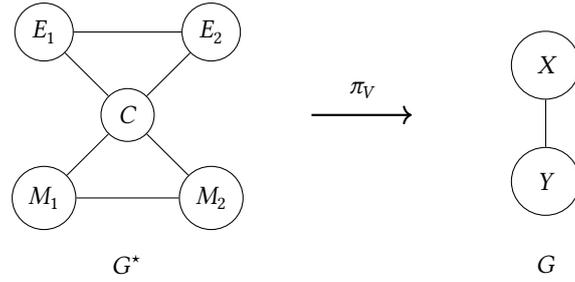


Figure 3: Ground-truth moral interaction graph G^* (left) and a coarse abstraction G (right) induced by aggregation map π_V with $X = \{E_1, E_2, C\}$ and $Y = \{M_1, M_2\}$.

Mechanism of Depth Reversal. With a short horizon, the planner mainly sees the immediate benefit of cutting the bridge between extremists and moderates: sanctioning C quickly reduces visible spillover, which appears strongly beneficial in the first few steps. However, sanctioning C also creates resentment at a highly connected node. That resentment spreads gradually through all of C 's links before fading, eventually affecting the entire network. A shallow evaluation captures the fast reduction in spillover but misses this slower, system-wide backlash. A deeper evaluation sees both effects, and once the delayed costs are included, targeted sanctioning becomes preferable.

Breadth Constraints

We now examine how abstraction alters the feasible intervention set and therefore the attainable moral value.

Medium abstraction (3 nodes). Let

$$\pi_V(E_1) = \pi_V(E_2) = X_E, \quad \pi_V(C) = X_C, \quad \pi_V(M_1) = \pi_V(M_2) = X_M.$$

This representation preserves the morally relevant distinction between extremists and the connector. Under this abstraction, the action $a^{(E)}$ remains admissible, and with sufficient depth the planner can recover the infinite-horizon optimum.

Coarse abstraction (2 nodes). Let

$$\pi_V(E_1) = \pi_V(E_2) = \pi_V(C) = X, \quad \pi_V(M_1) = \pi_V(M_2) = Y.$$

Admissible actions must satisfy

$$a_{E_1} = a_{E_2} = a_C,$$

so $a^{(E)} \notin \mathcal{A}_H(G_{\text{coarse}})$.

Under this abstraction (Figure 3), the planner cannot implement the infinite-horizon optimal action. Even with large depth, the best admissible intervention coincides with $a^{(EC)}$, whose infinite-horizon value is

$$M^*(a^{(EC)}, \omega) \approx -0.955,$$

strictly below the optimum

$$M^*(a^{(E)}, \omega) \approx -0.634.$$

The abstraction therefore induces an irreducible regret of approximately 0.322, independent of depth.

Regret and Moral Progress

Let $\rho^{(2)}(\omega, B) = (G^*, 2)$ and $\rho^{(10)}(\omega, B) = (G^*, 10)$, assuming the budget permits either allocation.

Using the infinite-horizon values above,

$$R(\omega; \rho^{(2)}, B) \approx 0.322, \quad R(\omega; \rho^{(10)}, B) = 0.$$

Under the coarse abstraction G_{coarse} , regret of the same magnitude persists even with large depth, since the optimal action is infeasible.

Takeaway. Limited depth hides delayed consequences. Limited breadth hides feasible interventions. Both forms of regret arise from constrained computation rather than disagreement about values. Moral progress corresponds to improving how computational resources are allocated across these two dimensions.