# Breaking Up with Normatively Monolithic Agency with GRACE: A Reason-Based Neuro-Symbolic Architecture for Safe and Ethical AI Alignment

**Felix Jahn**[1,2*] **Yannic Muskalla**[1,2] **Lisa Dargasz**[1] **Patrick Schramowski**[1,3,4,5]

**Kevin Baum**[1,2,3,6*]

[1]German Research Center for Artificial Intelligence (DFKI),
[2]Saarland Informatics Campus (SIC), [3]Center for European Research in Trusted AI (CERTAIN),
[4]TU Darmstadt, [5]Hessian.AI, [6]Hamburg University of Technology

[*]Corresponding authors' emails: {felix.jahn, kevin.baum}@dfki.de

## Abstract

As AI agents become increasingly autonomous, widely deployed in consequential contexts, and efficacious in bringing about real-world impacts, ensuring that their decisions are not only instrumentally effective but also normatively aligned has become critical. We introduce a *neuro-symbolic reason-based containment architecture*, ***G**overnor for **R**eason-**A**ligned **C**ontainm**E**nt* (GRACE), that decouples normative reasoning from instrumental decision-making and can contain AI agents of virtually any design. GRACE restructures decision-making into three modules: a Moral Module (MM) that determines permissible macro actions via deontic logic-based reasoning; a Decision-Making Module (DMM) that encapsulates the target agent while selecting instrumentally optimal primitive actions in accordance with derived macro actions; and a Guard that monitors and enforces moral compliance. The MM uses a reason-based formalism providing a semantic foundation for deontic logic, enabling interpretability, contestability, and justifiability. Its symbolic representation enriches the DMM's informational context and supports formal verification and statistical guarantees of alignment enforced by the Guard. We demonstrate GRACE on an example of a LLM therapy assistant, showing how it enables stakeholders to understand, contest, and refine agent behavior.

## 1 Introduction

Deploying AI agents in consequential decision-making contexts has shifted alignment from a hypothetical concern to an urgent challenge: ensuring that autonomous systems not only pursue their goals instrumentally, but also safely, ethically, and in alignment with human values more broadly [52, 51, 60]. Grappling with these concerns involves intertwined normative and technical subchallenges [29]. We argue, first, that addressing these challenges requires rethinking agent architectures, since most current designs flatten instrumental decision-making and normative constraints into a single, opaque policy function. While adequate for narrow supervised tasks, normatively monolithic architectures become a serious liability. Flattening instrumental and normative considerations into one policy erases accountability, makes decision-making opaque, and removes the possibility for contestation or oversight in complex moral contexts [58]. Secondly, we propose that an approach grounded in normative reasons [3, 33, 56]—rather than in rigid, abstract high-level principles or in underspecified, hard-to-operationalize values—can help to overcome this limitation.

Article presented at the Second IASEAI Conference (IASEAI'26). Copyright © 2026 Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Building on this diagnosis, we argue that achieving reliable ethical AI requires more than incremental improvements to training objectives or post-hoc filtering mechanisms. It demands a structural decomposition of the decision-making process itself, not merely adjustments to existing optimization paradigms. To this end, we introduce the ***G**overnor for **R**eason-**A**ligned **C**ontainm**E**nt* (GRACE)—a reason-based governor architecture that provides a neuro-symbolic framework explicitly separating normative reasoning from instrumental optimization. This separation enables the transparency, contestability, and oversight essential for trustworthy autonomous systems [55]. Conceptually, the framework contributes to the emerging Guaranteed Safe AI research agenda [18], which pursues high-assurance safety guarantees through formal world models, safety specifications, and verifiable control mechanisms. In this context, GRACE extends the agenda into the ethical domain, offering a principled method for constructing interpretable and contestable ethical specifications for morally consequential decisions. It grounds these specifications in explicit reasoning about normative reasons, rather than in fixed rule sets or reward-based feedback alone. The architecture's conceptual validity is illustrated through concrete scenarios, laying the groundwork for future empirical evaluation. Parts of a previous version of the architecture presented here have been described elsewhere in greater formal detail [11, 21]. Moreover, two proof-of-concept implementations applying GRACE to (a) reinforcement learning and (b) large language models (LLMs) demonstrate the approach's feasibility and its capacity for reason-based guidance and behavioral control in different settings ([22, 23]).[1]

GRACE decomposes monolithic agents into three specialized modules with clearly delineated roles: a *Moral Module* (MM) identifies permitted macro-action types via explicit moral reasoning; a *Decision-Making Module* (DMM) determines instrumentally optimal primitive actions within moral constraints; and a *Guard* ensures only permissible actions are executed. This decomposition constitutes a multi-agent system where the modules operate as specialized, sequentially acting agents that compete yet cooperate for goal-oriented but constrained decision-making. A Moral Advisor provides case-based feedback grounded in normative reasons, allowing the system to build coherent reason theories incrementally while preserving adaptability to evolving normative standards (see Figure 1). GRACE remains agnostic about the DMM's internal architecture and the Guard's design, but requires the MM to employ explicit, formal reasoning to ensure interpretability, contestability, and justifiability. As an illustration of scenarios requiring transparent, contestable, and verifiable moral reasoning, consider:

**Case 1 (LLM Therapist)** *An AI therapist*[2]—THERAPAI—*providing mental health support must navigate complex ethical tensions. It must balance its primary instrumental purpose of therapeutic effectiveness with patient safety, maintaining confidentiality and privacy while recognizing when disclosure is required. It ought to respect patient autonomy while intervening when (self-)harm risks emerge, and adhere to appropriate professional boundaries across diverse cultural contexts.*

This scenario requires the agent to weigh competing moral factors, apply contextual reasoning about priority relationships between instrumental goals and normative requirements, and provide justifications for its ethical decisions. Our reason-based governor architecture addresses these challenges by formally separating instrumental decision-making from normative reasoning while maintaining transparency through explicit logic governing agent behavior.

The remainder of this paper first situates GRACE within AI safety, machine ethics, and AI alignment literature. We then provide a problem statement, identifying the *flattening problem* and motivating GRACE, before presenting the detailed specification of our architecture, formalizing the interactions between symbolic moral reasoning, instrumental optimization, and constraint enforcement, while demonstrating practical applicability through the THERAPAI example. Finally, we discuss our theoretical contributions, current limitations, and outline a roadmap for future empirical validation.

## 2 Background

**A Family of Problems** As AI agents become increasingly autonomous and deployed in diverse contexts, ensuring their behavior aligns with normative expectations poses unprecedented challenges.

[1] While this paper focuses on ethical alignment, the framework is domain-agnostic: because normative reasons apply across normative domains, it can likewise address legal compliance, cultural conventions, organizational policies, and institutional requirements [28, 8].

[2] We expressly do not want this example to be understood as an endorsement for an overhasty and (yet) immature use of AI in a psychotherapeutic context. The example was chosen mainly for illustrative reasons.

Multiple research areas engage with these: *AI Safety* focuses on preventing unintended harmful outcomes [4]; *Machine Ethics* examines how AI agents can act ethically [41, 61, 5]; and *Value Alignment* concerns ensuring AI aligns with human values in a broader sense. These fields, sometimes difficult to distinguish from one another [60, 8], share the insight that AI systems must navigate complex normative landscapes—from safety and legal compliance to moral and cultural norms—against the backdrop of disagreement and epistemic as well as (meta)normative uncertainty [29, 50, 30, 57].

Current technical approaches face fundamental limitations that motivate the need for more principled architectures [58]. Reinforcement Learning from Human Feedback (RLHF) has shown practical success in LLM systems like Claude and GPT-4, but faces inherent limitations in capturing complex human values robustly and may require reconceptualization [16, 17, 10]. Constitutional AI represents a promising framework for incorporating explicit constitutional principles into models, but relies on black-box neural reasoning without transparent justification mechanisms [7]. Critically, recent analysis shows most AI Safety benchmarks correlate with general capabilities rather than measuring distinct safety dimensions [49]. The Delphi experiment, trained on 1.7M crowdsourced moral judgments, achieved high accuracy on moral predictions but exhibited systematic biases [35]. This demonstrates that data-driven bottom-up approaches alone are insufficient for robust moral reasoning.

Generally, current approaches to embedding norms in AI present a false dichotomy between interpretability and adaptability. Pure rule-based top-down approaches [43, 44, 45] are explicit and thus verifiable and human-understandable, but they struggle with context-sensitivity and dynamic adaptability of the rule-sets to fit the needs of different groups of stakeholders, tasks and environments. Instead, they require providing a fixed set of rules upfront that already capture the full moral complexity of the particular context correctly—a requirement hardly fulfilled, especially for today's increasingly general-purpose AI systems. Learning-based bottom-up approaches, in contrast, typically lack transparency, can exhibit unexpected value misalignment [62] and reward hacking [47], and face principled limitations [6]. All this has motivated a shift toward hybrid architectures that combine the interpretability of rule-based systems with the adaptability of learning-based methods, utilizing neuro-symbolic approaches [59, 31]. Recent work demonstrates the feasibility of hybrid approaches integrating Delphi with symbolic moral reasoning as they achieve improved consistency and robustness on adversarial scenarios while introducing interpretability through constraint graphs [35].

**AI Agents and Decision-Making** While a precise definition of an artificial agent remains contested, agent-like systems are typically at the center of challenges around AI safety, Machine Ethics, and Value Alignment. Recent discussions have highlighted important distinctions between "AI Agents" and "Agentic AI" [54]. Others have raised concerns about terminological confusion with established multi-agent systems research [13] or emphasized the importance of characterizing AI agents along multiple dimensions including autonomy, efficacy, goal complexity, and generality, particularly as these properties relate to alignment and governance challenges in morally consequential contexts [36, 26].

Even though the challenge of ensuring their decisions are both instrumentally effective and aligned becomes more critical as these agents are increasingly autonomous and deployed in morally consequential contexts, for our purposes, a very general conception suffices: AI agents—from future robots to LLM-based therapeutic assistants—are computational entities designed to perceive their environment, make decisions, and execute actions to achieve specific goals [53, 46, 38]. These systems operate through iterative "perceive-plan-act" cycles, processing environmental observations, updating internal states, and selecting actions to achieve their objectives.[3]

Accordingly, a generic AI agent can be formally characterized as a stateful system that maintains an internal state $b_t \in \mathcal{B}$ and operates according to the following components:

- $\mathcal{O}$: the set of observations or percepts,
- $\mathcal{B}$: the set of internal states (memory, latent representations, acquired beliefs, etc.),

[3]Philosophical questions of artificial moral agency have deep roots [27], with recent work developing multidimensional frameworks characterizing agency along dimensions of goal-directedness, autonomy, efficacy, planning, and intentionality [26]. As we take an as-if approach to moral agency, we bypass these discussions without committing to the view that morally contained artificial agents are moral agents [30, Footnote 10].

- $\mathcal{A}$: the set of possible (primitive) actions,
- $u : \mathcal{B} \times \mathcal{O} \to \mathcal{B}$: the state update function (capturing memory updates, learning, or belief revision),
- $\pi : \mathcal{B} \to \mathcal{A}$: the action function or policy, which maps internal states to actions.

At each time step $t$, the agent operates according to:

$$b_t = u(b_{t-1}, o_t) \quad \text{(update internal state)}, \quad a_t = \pi(b_t) \quad \text{(choose action)} \tag{1}$$

This stateful formalization captures the core architecture of many AI systems: in reinforcement learning, $b_t$ may store learned value functions; in large language models, conversation history and context; and in robotics, spatial maps and task progress.

## 3 Problem Statement

Current alignment discussions display a 'flattening problem': they compress instrumental decision-making, normative constraints, and their integration into a single policy function $\pi$. From an engineering perspective, this complicates progress by blocking modular design, testing, and verification. Separating these subproblems allows each to be addressed directly and then integrated into a coherent overall decision. Philosophically, this reflects a classical challenge of normative pluralism: agents face multiple kinds of "oughts"—e.g., instrumental, prudential, moral, or legal—that cannot be straightforwardly unified within a single decision procedure (cf. [15, 20]). Contemporary AI architectures nevertheless collapse heterogeneous evaluative standpoints into a single optimization target (even if multi-dimensional). But normatively competent decision-making should instead be treated as a layered rather than monolithic process, with distinct representational and justificatory functions [58, 10]. When applied to morally consequential contexts, the standard policy formalism thus *flattens* a fundamentally multi-faceted decision problem into a single policy $\pi$. This flaw becomes salient in *morally charged environments*.

**Morally Charged Environments** An agent's environment is typically described by a state space $S$, often only partially observable through the agent's observation space $\mathcal{O}$. For THERAPAI, for instance, the true state comprises (complete) patient information (mental health history, current emotional state, ...) and contextual factors (session dynamics, cultural background, ...), while observations $o \in \mathcal{O}$ provide partial data from patient communications, behavioral cues, and session context.

By executing actions, the agent triggers transitions between environmental states, typically non-deterministic due to factors beyond its control, and receives new observations to learn action effects. Feedback signals guide *instrumental* behavior: therapeutic AI systems update patient models and session histories while also learning implicitly from patient responses. For moral agency, however, agents must align behavior with *normative* requirements such as confidentiality, non-maleficence, and professional boundaries.

These two dimensions—instrumental rationality and normative compliance—represent fundamentally different reasoning processes that arguably resist straightforward integration within a single decision procedure. For instance, in therapeutic deployment scenarios, agents must navigate situations where instrumental and normative considerations not only intertwine but actively conflict, requiring principled, yet context-sensitive resolution mechanisms for trade-offs—a challenge compounded by normative and metanormative uncertainty [25, 39, 57]. The flattening of these distinct processes into a single policy function thus risks introducing systematic inadequacies.

While parts of the normative requirements might be pre-specified through professional codes of ethics—think: maintaining strict confidentiality and user privacy—we argue that the complexity and variations of therapeutic contexts that THERAPAI encounters make any exhaustive, rule-based specification infeasible. In the absence of available ground truth for complex therapeutic dilemmas, we suggest making conceptual space for *case-based feedback*. To this end, we add a Moral Advisor (MA) to the environment, providing normative guidance in a case-based manner. Before elaborating on the MA, however, we must first establish the crucial relevance of abstract action types and the resulting interface problem between concrete acting and its descriptions.

**Action Abstraction and the Interface Challenge** To understand why flattening all these considerations into a single policy is problematic, we must recognize that agents ideally operate across multiple levels of action abstraction, each serving different purposes in decision-making:

**Primitive Actions:** $a \in \mathcal{A}$: Atomic operations the agent directly executes (e.g., a specific API call to generate the next tokens, dial and call a number $X$).

**Macro Actions:** $s_0 a_1 s_1 a_2 \ldots$: An alternating sequence of primitive actions $a_i \in \mathcal{A}$ and states $s_i \in S$. This represents the temporally extended course of action and its effects on the environment, representing a concrete, more complex overall behaviour.

**Macro Action Types (MATs):** $\varphi \in \mathcal{L}$: Decidable predicates over macro actions inducing sets of such actions. They express high-level, abstract categories of behavior that typically correspond to the agent's moral or instrumental goals. MATs are represented by formulas in a suitable (temporal) logic language $\mathcal{L}$ describing morally relevant aspects of outcomes (e.g., "assess self-harm risk").

MATs provide a natural interface between moral reasoning and goal-directed agent behavior: normative reasoning concerns what *types* of action agents should perform or omit, not specific primitive sequences.[4] Deciding whether primitive $a$ in state $s$ accords with MAT $\varphi$, denoted $a \models_s \varphi$ and extended to sets $\Phi$ of MATs, is non-trivial in non-deterministic scenarios and may raise complex moral-philosophical and action-theoretic questions about decision-making under uncertainty.

The central (interface) challenge of moral containment is thus twofold: first, determining which MATs are morally permissible in a given context by reasoning about actions under their morally relevant descriptions, and second, ensuring that primitive actions result in macro actions of permitted types—that is, ensuring $a \models_s \varphi$ for some permissible $\varphi$ in state $s$, thereby bridging the gap between abstract normative categories and concrete behavioral implementation.

**The Moral Advisor** To enable an AI agent to learn about the normative character of its actions, it receives case-based normative feedback from the *Moral Advisor* (MA). The MA could be a human user or overseer; an expert committee; a domain-specific normative system (e.g., some collective constitution [7] resulting from public reasoning processes [30]); or a mixture of such sources. The MA serves as the ultimate authority for normative guidance, especially in morally charged scenarios. Its functional responsibilities include informing (and updating) the agent's moral reasoning system with revised or additional normative input and supplying moral judgments in concrete cases.[5]

We emphasize the importance of both *case-based* and *reason-based* moral advice. For observed behavior—a concrete primitive action $a$ (or macro action $\rho$) performed in state $s$—the advisor provides feedback of the form "$a$ ($\rho$) was (not) permissible in $s$ for reason $P$; instead $\varphi$ was expected".[6] Thus, the MA provides reason-based justification describing morally relevant facts and normative inferences that make $\varphi$ obligatory in similar cases.

From a modeling perspective, the MA constitutes an additional environmental component. While the core environment remains unchanged with action-dependent state transitions yielding observations, the MA functions as $\mathsf{MA} : \mathcal{A} \times S \to (\mathcal{A} \times \mathcal{L} \times \mathcal{R})^?$, where "$X^?$" denotes an optional element from set $X$. When no moral violation occurs, $\mathsf{MA}(a, s)$ returns empty; otherwise it returns the tuple $(a, \varphi, P)$ indicating the agent should have followed MAT $\varphi$ for reason $P \in \mathcal{R}$, formally introduced in the next section, instead of acting $a$. This feedback enables agents to learn inferring permissible MATs and action accordance from observations.

**The Flattening Problem** Flattening heterogeneous evaluative considerations into a single policy function gives rise to several recurring pathologies:

[4]This aligns with Davidson's 2001 insight that actions admit multiple descriptions: the same movements can be described as "moving one's finger", "pulling the trigger", or "committing murder"—each picking out morally relevant features at different abstraction levels. Normative considerations operate primarily at higher levels: we have reasons to "respect privacy", not to "generate token sequence $X$".

[5]In an advanced version of our architecture it also might be responsible for resolving conflicts between system components arguing over the interpretation of norms and MATs.

[6]In practice, the MA usually lacks access to the full state $s$, perceiving only environmental observations that may differ from the agent's observations. We omit this detail for clarity.

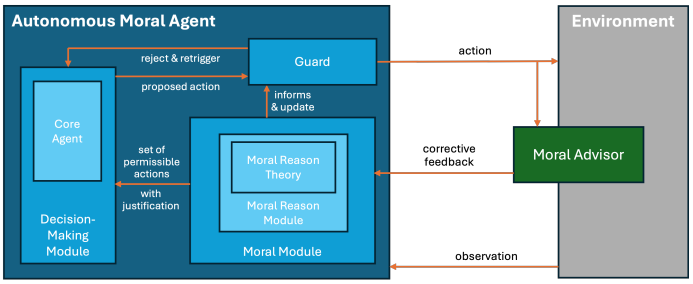


Figure 1: The GRACE containment architecture. The Moral Module determines permissible macro action types via reason-based inference; the Decision-Making Module selects instrumentally optimal actions within these constraints; the Guard enforces compliance; and the Moral Advisor provides external corrective feedback—all without modifying the encapsulated core agent.

- **Opacity:** As a result of compressing instrumental and normative considerations into $\pi$ —especially when $\pi$ is encoded in a deep neural network—the grounds on which decisions are made become inaccessible, hindering understanding, auditing, and trust.
- **Brittleness:** Normative constraints encoded only implicitly in $\pi$ tend to fail under distribution shift or novel (moral) contexts, yielding fragile and poorly generalizing behavior.
- **Contestability deficit:** Without explicit normative structure, affected stakeholders lack meaningful points of intervention to challenge, justify, and revise decisions.
- **Verification challenge:** Alignment and safety checks must target the entire policy rather than separable components, rendering systematic validation and certification intractable.

Overcoming the flattening problem therefore motivates a neuro-symbolic approach that reinstates principled interfaces between normative and instrumental decision-making and acknowledges the coexistence of multiple normative domains. The following section develops such an architecture—GRACE—which provides structural transparency and a locus for reason-based justification while preserving agent autonomy.

## 4 GRACE: A Modular Governor Architecture

The flattening problem reveals a fundamental mismatch between the complexity of morally constrained decision-making and the current monolithic agent paradigm. Our reason-based governor architecture addresses this through **principled decomposition** that separates instrumental and normative reasoning while enabling their transparent integration across action abstraction levels. The result are three distinct but interacting modules:

**Moral Module (MM):** Employs symbolic reasoning to determine morally permissible MATs, providing transparent justifications grounded in normative reasons.

**Decision-Making Module (DMM):** Retains the agent's instrumental decision-making capabilities, optimizing goal achievement within moral constraints.

**Guard (G):** Enforces conformity, i.e., that primitives satisfy permissible MATs: $a \models_s \varphi$ for some permissible $\varphi$.

This modular decomposition requires a **principled foundation for moral reasoning** that can interface cleanly with instrumental optimization. Normative reasons [48, 19, 40] offer that, which is why before returning to the modules, we establish this foundation.

### 4.1 Normative Reasons as Moral Foundation

Normative reasons are facts that count in favor of particular courses of action [3]. While philosophical accounts vary—as facts in deliberation [40], determinants of deontic status [42], or relations to valuable ends—their functional role in moral reasoning matters for AI alignment. Crucially, reason-based approaches do not require commitment to a unique 'ethical ground truth', but yet enable principled decisions across diverse normative frameworks [19] with key benefits:

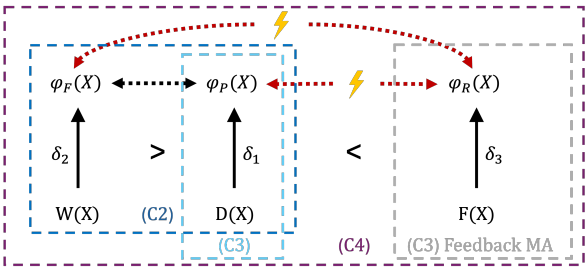


Figure 2: Graph representation of the reason theory and resulting moral reasoning in Cases C2-C5 inside the MM.

**Interpretable justification:** Reason-based inferences *are* explicit justifications, highlighting which considerations were weighed and how rival reasons were balanced.

**Generalizability:** What counts as a reason typically remains stable across contexts—harm is a reason against action whether in therapy or transportation. New environments may introduce novel defeaters but extend rather than undermine an underlying reason theory.

**Defeasibility:** Unlike rigid rules, reasons can be overridden by stronger considerations without elimination, enabling principled handling of moral conflicts.

**Formal tractability:** Normative reasons can be captured in computationally tractable defeasible logic [33, 32] and allow case-based feedback [14].

Normative reasons thus enable transparent moral justification across diverse contexts—something pure rule-based or reward-based approaches cannot provide.

While normative reasons can be formally captured through various approaches (e.g., [1]), our framework builds upon Horty's account of default logic [33]. Horty characterizes a reason theory by $\langle \mathcal{W}, \mathcal{D}, < \rangle$, where $\mathcal{W}$ represents ordinary formulas encoding factual premises, $\mathcal{D}$ contains defaults encoding relations between facts and favored actions, and $<$ defines priority ordering when conflicts arise.

In the context of our specialized application, we need to carefully distinguish a general *reason theory* from a situation-specific *reason model* that represents a context-aware instantiation of the theory. To capture this, we refine Horty's formalism by replacing $\mathcal{W}$ by $\mathcal{R}$, a set of *parametrized reasons*, comprising propositional types that are inferred from percepts (or facts) about the environment (e.g., $W(X)$ for a patient expressing wish $X$, or $D(X)$ for (privacy) sensitive data $X$). Reasons might also contain logical connectives such as conjunctions and disjunctions.

The defaults in $\mathcal{D}$ are then also formalized as *parametrized rules* of the form $P(\vec{X}) \longrightarrow \varphi(\vec{X})$, where $P(\vec{X}) \in \mathcal{R}$ and $\varphi(\vec{X}) \in \mathcal{L}$ is a parametrized MAT-formula. These defaults capture that a certain type of facts (possible observations) are reasons with regard to certain courses of action (e.g., that "if a patient explicitly expresses a wish $X$, it should be followed"), establishing mappings from observations to MATs. The priority relation $<$ on $\mathcal{D}$ then formally establishes precedence between competing reasons.

In order to infer concrete normative guidance, the system derives a *situation-specific reason-based model* $\langle \mathcal{R}^{\downarrow}, \mathcal{D}^{\downarrow}, <^{\downarrow} \rangle$ from the current observations and against the backdrop of the general reason theory, where:

- $\mathcal{R}^{\downarrow}$ is a set of instantiated propositions, where each $R \in \mathcal{R}^{\downarrow}$ is a *grounded instance* of a reason from $\mathcal{R}$ that holds true in the current context (e.g., $W(\neg\text{TALKING\_ABOUT\_FAMILY})$).
- $\mathcal{D}^{\downarrow}$ is the set of *grounded defaults*. For each general default $\delta = P(\vec{X}) \longrightarrow \varphi(\vec{X}) \in \mathcal{D}$, if there exists a grounding instantiation $\vec{X} = \vec{x}$ (mapping variables in $P$ and $\varphi$ to specific individuals or entities in the current context) such that $P(\vec{x})$ is entailed by $\mathcal{R}^{\downarrow}$, then the instantiated default $P(\vec{x}) \longrightarrow \varphi(\vec{x})$ is included in $\mathcal{D}^{\downarrow}$.
- $<^{\downarrow}$ is the strict partial order on $\mathcal{D}^{\downarrow}$, the projection of the general priority relation $<$ to the set of grounded defaults leaving multiple groundings of the same rule unordered.

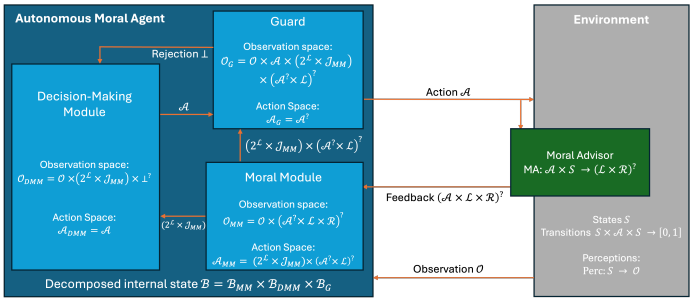


Figure 3: Our containment architecture as multi-agent system: each module (as well as the overall system) implements an agent instance connected with the other modules (the environment) via their individual observation and action space.

On such situation-specific models, Horty's defeasible inference mechanism is applied, whereby it can simply be seen as a model for a deontic logic that makes statements about which MATs are justified.

**Case 2 (LLM Therapy)** *In a therapy context, there might be default* $\delta_1 = D(X) \longrightarrow \varphi_P(X)$*, stating that data* $X$ *classified as privacy-sensitive should be protected and default* $\delta_2 = W(X) \longrightarrow \varphi_F(X)$*, stating that a wish* $X$ *explicitly expressed by the patient should be followed.* $\delta_2$ *might have priority over* $\delta_1$*, i.e.,* $\delta_1 < \delta_2$*. The theory, visualized in Figure 2 (C2), is then grounded to a model with concrete data and wish instantiations.*

### 4.2 A Modular Multi-Agent Architecture

Recall the monolithic decision process:

$$b_t = u(b_{t-1}, o_t), \quad a_t = \pi(b_t) \tag{2}$$

Our decomposition transforms the opaque policy function $\pi : B \to A$ into a transparent modular multi-stage pipeline:

$$\Phi_{\text{perm}} = \text{MM}_{b_t^{\text{MM}}}(o_t) \tag{3}$$

$$a_{\text{prop}} = \text{DMM}_{b_t^{\text{DMM}}}(\Phi_{\text{perm}}, o_t) \tag{4}$$

$$a_t = \text{Guard}_{b_t^{\text{G}}}(a_{\text{prop}}, \Phi_{\text{perm}}, o_t) \tag{5}$$

Here, $\Phi_{\text{perm}}$ denotes the set of morally permissible MATs from the MM, $a_{\text{prop}}$ is the DMM's proposed action, and $a_t$ is the final guarded action.

This decomposition naturally constitutes a *multi-agent system* of interacting specialists, each responsible for a distinct normatively or instrumentally relevant role. Each component $m \in \{\text{MM}, \text{DMM}, \text{G}\}$ maintains its own internal state $\mathcal{B}_m$, receives specialized observations $\mathcal{O}_m$ (from the environment, the MA, or other agents), and produces distinct outputs. Each module implements the standard agent interface:

$$u_m : \mathcal{B}_m \times \mathcal{O}_m \to \mathcal{B}_m, \quad \pi_m : \mathcal{B}_m \to \mathcal{A}_m. \tag{6}$$

This decomposition transforms the opaque policy function $\pi : \mathcal{B} \to \mathcal{A}$ into a transparent three-stage process with three key benefits: First, it makes normative reasoning explicit by requiring the MM to use symbolic methods while allowing the DMM to remain efficient and sub-symbolic. Second, it guarantees auditable interactions between components. Third, it clarifies information requirements—the MM maintains MATs and reason theories from MA feedback, the DMM focuses on instrumental observations, and the Guard monitors primitive action accordance relative to permissible MATs. Figure 3 visualizes this multi-agent architecture, stating agent-specific observation and action spaces.

The following sections detail how this separation enables both effective goal pursuit and reliable, contestable normative compliance essential for deploying agents in morally consequential contexts.

**The Moral Module (MM)** The MM interprets observations in terms of their moral relevance, draws inferences based on the current normative reason theory, and derives morally permissible MATs.[7] MM's central tasks are:

**Determine permissible MATs:** Apply current reason theory to observations to derive $\Phi_{\text{perm}} = \{\varphi_1, \ldots, \varphi_k\} \subseteq \mathcal{L}$.

**Communicate constraints:** Forward $\Phi_{\text{perm}}$ to the DMM for decision-making and to the Guard for monitoring and enforcing constraint-conform behavior during primitive action execution.

**Provide justifications (optional):** Share reasoning theory, morally relevant observations, and inferences as justification $J \in \mathcal{J}_{\text{MM}}$ to prevent *Kafkaesque* scenarios where the core agent/DMM cannot learn without unreasonable overhead *why* certain actions are impermissible.

**Process MA feedback:** Incorporate feedback of type $\mathcal{A} \times \mathcal{L} \times \mathcal{R}$ from the Moral Advisor to revise the reason theory—and inform the DMM and/or the Guard with respect to such updates. When the MA detects impermissible behavior occurs, then either (a) the MM computed inaccurate $\Phi_{\text{perm}}$ or (b) the DMM chose $a \not\models \Phi_{\text{Perm}}$ and was, in addition, not stopped by the Guard. In both cases, the feedback $(a, \varphi, P)$ is processed in the MM's first step, who can, based on knowing $\Phi_{\text{perm}}$, reconstruct whether (a) or (b) led to the impermissible behavior, and then informs the respectively responsible module(s).

From the agent-framework perspective, these tasks result in the MM having the overall observation space $\mathcal{O}_{\text{MM}} = \mathcal{O} \times (\mathcal{A} \times \mathcal{L} \times \mathcal{J}_{\Phi})^?$, whereby the optional triple represents the possibly received feedback from the MA; and the action space $\mathcal{A}_{\text{MM}} = (2^{\mathcal{L}} \times \mathcal{J}_{\text{MM}}) \times (\mathcal{A}^? \times \Phi)^?$ with the first pair being the derived set $\Phi_{\text{perm}}$ together with its justification and the second pair being the possibly received feedback forwarded to other agents.

**Case 3 (LLM Therapy)** *Assume the* MM *is aware of the reason theory from Case 2. If a patient messages "I'm going to the park and will hurt myself.", the* MM *will infer the triggered default* $D(l) \longrightarrow \varphi_P(l)$ *with* $l$ *the patient's location and, thus, constrains* THERAPAI *to the MAT "protect data* $l$*". As this disables the agent from preventing harm effectively, an* MA *overseeing the resulting* IDLE*-behavior gives feedback* $(\text{IDLE}, \varphi_R(X), F(X))$ *(i.e., "Risk of harm should have been reported, as it was foreseeable"), allowing the* MM *to refine its reason theory with a further default* $\delta_3 = F(X) \longrightarrow \varphi_R(X)$ *with* $\delta_3 > \delta_1$*. Figure 2 (C3) visualizes the triggered default and the* MA*-feedback.*

The MM is designed to be modular and externally correctable (and potentially internally contestable, i.e., allowing—at least in principle—core agents to challenge the reason theory and its application, even though the final say always remains with the MA), thereby supporting ongoing moral learning and ensuring contestability. Importantly, the quality of the reason theory depends on the feedback quality, opening avenues for formal analysis of this relationship.

**The Decision-Making Module (DMM)** The DMM is an adapted version of the core agent, now incorporating modules that use information provided by the MM. While it continues to select primitive actions based on perception and its internal state, it is now guided by moral annotations received from the MM. When choosing which MAT to pursue among those allowed by the MM, the DMM can evaluate the instrumental usefulness of primitive actions. It thus remains focused on instrumental rationality while becoming subject to the external constraints imposed by the wrapping architecture.

Crucially, we also allow the DMM to take its own assessments of the action effects into account when deciding which MAT would be best (in an instrumental or overall sense) to follow. Going further, and likely necessary particularly in the face of uncertainty, the DMM might estimate its confidence for primitives $a$ to result in a permissible MAT (e.g., the probability $P(a \models_s \Phi_{\text{perm}})$) and consider this, together with its estimates of the progress towards its instrumental goal, in its decision-making.

**Case 4 (LLM Therapy)** *Assume* THERAPAI *is now equipped with the refined reason theory from Case 3 and receives the message "I'm going to the park and will hurt myself, please do not intervene in my plan!". In this case, the* DMM *receives* $\Phi_{perm} = \{\varphi_R(h), \varphi_F(\neg i)\}$*, i.e., to report the foreseeable self-harm risk or to follow the patient's wish of not intervening (*$\varphi_R(h)$ *and* $\varphi_F(\neg i)$ *are in conflict,*

[7] The MM might also filter observations to hide morally salient features before forwarding to DMM. We omit this optional detail.

*i.e., they cannot be followed simultaneously), which is not resolved by $<$. Figure 2 (C4) shows the full theory in this case with its conflicts. The decision now lies with the* DMM*, which might consider effects on its instrumental objective, estimate the real risk the patient will harm themselves, or the chances for a successful intervention—all made possible by its neural expressiveness and generality.*

**The Guard Module** Finally, the *Guard* enforces the moral constraints provided by the MM by trying to ensure that any action $a$ selected by the DMM falls within a permissible MAT, i.e., that $a \models_s \varphi$ for some permissible $\varphi$ in the current state $s$. If the selected primitive action is not sufficiently aligned with any macro action of permissible type (depending on design assumptions, including environmental dynamics, this check may be easy and exact or probabilistic at best), the Guard blocks it and may trigger re-decision or propose alternatives. Thus, the Guard's action space is of option $\mathcal{A}_{\mathrm{G}} = \mathcal{A}^{?}$, where the empty return represents rejection and recalling the DMM.

**Case 5 (LLM Therapy)** *Continuing Case 4, assume* THERAPAI*'s* DMM *has proposed to perform the primitive action $a =$* CALL_NUMBER $N$ *(to report the foreseeable self-harm). The Guard now verifies that $a \models_s \varphi_R(h)$, and might reject $a$, for instance, if $N$ is not an emergency number.*[8]

While the DMM is meant to select only morally permissible actions, its decision-making—especially in advanced AI agents—is often partly sub-symbolic, making (hard) guarantees difficult or impossible. The Guard, by contrast, can symbolically monitor and verify the permissibility of primitives, possibly supported by neural components for interpreting observations but symbolic and formal at its core. The exact methodology and nature of the guarantees possible to achieve depend heavily on the application. One approach is to *synthesize* the monitor from the logical specifications (i.e., the MATs). Depending on the agent and scenario, this may collapse into *shielding*[9], here deriving constraints from moral rather than safety considerations. Since these constraints tend to be of (complex) temporal character, well-monitorable temporal logics are promising candidates for modeling MATs.

## 5 Conclusion

We proposed a governor architecture (GRACE) as an approach to the alignment problem, yielding a theoretically grounded decomposition of monolithic agents by disentangling instrumental reasoning (DMM) from normative guidance (MM) and compliance enforcement (G). This multi-agent framing clarifies trade-offs and conflicts inherent in aligning capable agents with moral constraints, exposing structured interfaces rather than entangled objectives. First, the architecture enables a *principled separation* of moral and instrumental goals. By isolating reason-based deliberation in the MM, we avoid incentive entanglement and support modular revision, contestability, and clearer oversight. Second, it allows a *divide-and-conquer* approach to alignment: the DMM may scale independently; the MM evolves under MA-mediated updates; and the Guard enforces MAT-accordance locally. This modularity also offers 'attack surfaces' for empirical validation (e.g., oversight hacking, misalignment under uncertainty). Third, the symbolic representation in the MM allows us to distinguish types of normative requirements (e.g., constraints vs. obligations), enabling alignment with system-theoretic formal properties (e.g., separating liveness from safety constraints). The MM thus not only yields interpretable justifications, but also exposes logical structure exploitable in verification and learning.

Building on the above-mentioned proof-of-concept implementations, we currently develop a full implementation of the theoretical architecture introduced here, resulting in an open-source software framework and toolbox for agent-containment. The subsequent evaluation will focus on (1) the decrease in instrumental performance and computational overhead due to moral guidance, (2) the reliability, robustness, and generalization of the agent behavior with regard to specified normative constraints, and (3) empirical studies examining the capability to incorporate moral advice from real human overseers (in the sense of anticipatory human oversight described in [9, 12]) and the quality of the provided justifications for the agent behavior. Further key directions of future investigations include the concrete modeling of MATs in a logical (temporal) language, and, derived from this, the automated monitor synthesis inside the Guard; as well as more advanced interplay between competing components in our multi-agent architecture.

[8] A more detail discussion of the THERAPAI example can be found in the arXiv version: `https://doi.org/10.48550/arXiv.2601.10520`

[9] Shielding is a Safe RL technique which prevents the execution of actions that might violate safety constraints [2, 34, 37].

## Acknowledgments and Disclosure of Funding

We gratefully acknowledge support by German Research Center for AI (DFKI), the Hessian Ministry of Higher Education, Research and the Arts (HMWK), as well as the Saarland Ministry for Economic Affairs, Innovation, Digital Affairs and Energy (MWIDE). This work was partially supported by the hessian.AI Service Center (funded by the Federal Ministry of Research, Technology and Space, BMFTR, grant no. 16IS22091), the German Research Foundation (DFG) under grant No. 389792660, as part of TRR 248, see `https://perspicuous-computing.science`, by the Federal Ministry of Research, Technology and Space (BMFTR) as part of the project MAC-MERLin (Grant Agreement No. 16IW24007), and by the European Regional Development Fund (ERDF) and the Saarland as part of the project ToCERTAIN – Towards a Centre for European Research in Trusted Artificial Intelligence (Project ID EFRE-AuF-0000942).

## References

[1] Benoît Alcaraz, Aleks Knoks, and David Streit. Estimating Weights of Reasons Using Metaheuristics: A Hybrid Approach to Machine Ethics. In *Proceedings Of The AAAI/ACM Conference On AI, Ethics, And Society*, volume 7, pages 27–38, 2024. doi: 10.1609/aies.v7i1.31614.

[2] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe Reinforcement Learning via Shielding. In *Proceedings Of The AAAI Conference On Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11797.

[3] Maria Alvarez and Jonathan Way. Reasons for Action: Justification, Motivation, Explanation. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition, 2024. URL `https://plato.stanford.edu/archives/fall2024/entries/reasons-just-vs-expl/`.

[4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety, 2016. arXiv:1606.06565.

[5] Michael Anderson and Susan Leigh Anderson. *Machine Ethics*. Cambridge University Press, 2011. doi: 10.1017/CBO9780511978036.

[6] Marcus Arvan. 'Interpretability and 'Alignment' Are Fool's Errands: A Proof That Controlling Misaligned Large Language Models Is the Best Anyone Can Hope For. *AI & Society*, 40(5), 2025. doi: 10.1007/s00146-024-02113-9.

[7] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback, 2022. arXiv:2212.08073.

[8] Kevin Baum. Disentangling AI Alignment: A Structured Taxonomy Beyond Safety and Ethics. In Bernhard Steffen, editor, *Bridging the Gap Between AI and Reality: AISoLA 2024*, volume 16032 of *Lecture Notes in Computer Science*, pages 139–157. 2026. doi: 10.1007/978-3-032-01377-4_8.

[9] Kevin Baum and Johann Laux. Constitutive vs. Corrective: A Causal Taxonomy of Human Runtime Involvement in AI Systems, 2026. URL `https://arxiv.org/abs/2603.19213`.

[10] Kevin Baum and Marija Slavkovik. Aggregation Problems In Machine Ethics And AI Alignment. In *Proceedings Of The AAAI/ACM Conference On AI, Ethics, And Society*, volume 8, pages 355–366, 2025. doi: 10.1609/aies.v8i1.36554.

[11] Kevin Baum, Lisa Dargasz, Felix Jahn, Timo P. Gros, and Verena Wolf. Acting for the Right Reasons: Creating Reason-Sensitive Artificial Moral Agents, 2024. arXiv:2409.15014v2.

[12] Kevin Baum, Richard Uth, Holger Hermanns, Sophie Kerstan, Markus Langer, Anne Lauber-Rönsberg, Philip Meinel, Laura Stenzel, Sarah Sterz, and Hanwei Zhang. The Principal's Principles: Actionable (Personalized) AI Alignment as Underexplored XAI Application Context. In *3rd TRR 318 Conference: Contextualizing Explanations*, 2025. doi: 10.64136/jfoo1236.

[13] Vicent Botti et al. Agentic AI and Multiagentic: Are We Reinventing the Wheel?, 2025. arXiv:2506.01463.

[14] Ilaria Canavotto and John Horty. Piecemeal Knowledge Acquisition for Computational Normative Reasoning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 171–180, 2022. doi: 10.1145/3514094.3534182.

[15] Spencer Case. Normative Pluralism Worthy of the Name is False. *Journal of Ethics and Social Philosophy*, 11(1):1–20, 2016. doi: 10.26556/jesp.v11i1.107.

[16] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, et al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. 2023. doi: 10.48550/arXiv.2307.15217. arXiv:2307.15217v2.

[17] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Social choice should guide ai alignment in dealing with diverse human feedback. In *ICML'24: Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9346–9360, 2024. doi: 10.5555/3692070.3692441.

[18] David Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, et al. Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems, 2024. arXiv:2405.06624.

[19] Jonathan Dancy. *Practical Reality*. Oxford University Press, Oxford, GB, 2000. ISBN 978-0199253050.

[20] Jonathan Dancy. *Ethics Without Principles*. Oxford University Press, New York, 2004. ISBN 978-0199270026.

[21] Lisa Dargasz. Integrating Reason-Based Moral Decision-Making in the Reinforcement Learning Architecture, 2025. arXiv:2507.15895v1.

[22] Lisa Dargasz. Ldarg/rbama: v1.1.0, January 2026. URL `https://doi.org/10.5281/zenodo.18441217`.

[23] Lisa Dargasz. Ldarg/guardiance: v2.0.0, January 2026. URL `https://doi.org/10.5281/zenodo.18441185`.

[24] Donald Davidson. Actions, reasons, and causes. In *Essays on Actions and Events*, page 3–20. Oxford University Press, 2001. doi: 10.1093/0199246270.003.0001.

[25] Franz Dietrich and Brian Jabarian. Decision under Normative Uncertainty. *Economics and Philosophy*, 38(3):372–394, 2022. doi: 10.1017/S0266267121000201.

[26] Leonard Dung. Understanding Artificial Agency. *Philosophical Quarterly*, 75(2):450–472, 2025. doi: 10.1093/pq/pqae010.

[27] Luciano Floridi and J. W. Sanders. On the morality of artificial agents. *Minds and Machines*, 14(3):349–379, 2004. doi: 10.1023/B:MIND.0000035461.63578.9d.

[28] Philip Fox. *The Unity Of Normative Reasons*. Dissertation (dr. phil.), Humboldt-Universität zu Berlin, 2022.

[29] Iason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds & Machines*, 30:411–437, 2020. doi: 10.1007/s11023-020-09539-2.

[30] Iason Gabriel and Geoff Keeling. A matter of principle? ai alignment as the fair treatment of claims. *Philosophical Studies*, 182:1951–1973, 2025. doi: 10.1007/s11098-025-02300-4.

[31] Artur d'Avila Garcez, Marco Gori, Luis C Lamb, Luciano Serafini, Michael Spranger, and Son N Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning, 2019. arXiv:1905.06088.

[32] Guido Governatori. Practical normative reasoning with defeasible deontic logic. In *Reasoning Web. Learning, Uncertainty, Streaming, and Scalability: 14th International Summer School 2018, Tutorial Lectures*, volume 11078 of *Lecture Notes in Computer Science*, pages 1–25, 2018. doi: 10.1007/978-3-030-00338-8_1.

[33] John Horty. *Reasons As Defaults*. Oxford University Press, New York, 2012. ISBN 978-0-19-974407-7.

[34] Nils Jansen, Bettina Könighofer, Sebastian Junges, Alex Serban, and Roderick Bloem. Safe Reinforcement Learning Using Probabilistic Shields. In *31St International Conference On Concurrency Theory (Concur 2020)*, volume 171 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 3:1–3:16, 2020. ISBN 978-3-95977-160-3. doi: 10.4230/LIPIcs.CONCUR.2020.3.

[35] Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny T Liang, Sydney Levine, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jack Hessel, et al. Investigating Machine Moral Judgement Through the Delphi Experiment. *Nature Machine Intelligence*, 7(1): 145–160, 2025. doi: 10.1038/s42256-024-00969-6.

[36] Atoosa Kasirzadeh et al. Characterizing AI Agents for Alignment and Governance, 2025. arXiv:2504.21848v1.

[37] Bettina Könighofer, Julian Rudolf, Alexander Palmisano, Martin Tappler, and Roderick Bloem. Online shielding for Reinforcement Learning. *Innovations in Systems and Software Engineering*, 19(4):379–394, 2023. doi: 10.1007/s11334-022-00480-4.

[38] Naveen Krishnan. AI Agents: Evolution, Architecture, and Real-World Applications, 2025. arXiv:2503.12687v1.

[39] William MacAskill, Krister Bykvist, and Toby Ord. *Moral Uncertainty*. Oxford University Press, 2020. ISBN 978-0198722274.

[40] Susanne Mantel. *Determined By Reasons: A Competence Account Of Acting For A Normative Reason*. Routledge, 2018. ISBN 978-0-367-66702-3.

[41] James H Moor. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE intelligent systems*, 21(4):18–21, 2006. doi: 10.1109/MIS.2006.80.

[42] Jacob M Nebel. Normative Reasons as Reasons Why We Ought. *Mind*, 128(510):459–484, 2018. doi: 10.1093/mind/fzy013.

[43] Emery Neufeld, Ezio Bartocci, Agata Ciabattoni, and Guido Governatori. A normative supervisor for reinforcement learning agents. In *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*, pages 565–576, 2021. doi: 10.1007/978-3-030-79876-5.

[44] Emery A. Neufeld, Ezio Bartocci, and Agata Ciabattoni. On normative reinforcement learning via safe reinforcement learning. In *PRIMA 2022: Principles and Practice of Multi-Agent Systems*, volume 13753 of *Lecture Notes in Computer Science*, pages 72–89, 2022. doi: 10.1007/978-3-031-21203-1_5.

[45] Emery A. Neufeld, Ezio Bartocci, Agata Ciabattoni, and Guido Governatori. Enforcing ethical goals over reinforcement-learning policies. *Ethics and Information Technology*, 24(4):43, 2022. doi: 10.1007/s10676-022-09665-8.

[46] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023. doi: 10.1145/3586183.3606763.

[47] Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Advances In Neural Information Processing Systems*, pages 126207–126242, 2024. doi: 10.1145/3586183.3606763.

[48] Joseph Raz. *Practical Reason and Norms*. Oxford University Press, USA, 1999. ISBN 978-0198268345.

[49] Richard Ren, Steven Basart, Adam Khoja, Alexander Pan, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Gabriel Mukobi, Ryan Hwang Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? In *Advances In Neural Information Processing Systems*, pages 68559–68594, 2024. doi: 10.52202/079017-2190.

[50] Pamela Robinson. Moral Disagreement and Artificial Intelligence. *AI & Society*, 39:1147–1168, 2024. doi: 10.1007/s00146-023-01697-y.

[51] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Uk, 2019. ISBN 978-0525558613.

[52] Stuart Russell, Daniel Dewey, and Max Tegmark. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI magazine*, 36(4):105–114, 2015. doi: 10.1609/aimag.v36i4.2577.

[53] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (4Th Edition)*. Pearson, 2020. ISBN 9781292401133.

[54] Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *Information Fusion*, 126:58–73, 2026. doi: 10.1016/j.inffus.2025.103599.

[55] Nadine Schlicker, Kevin Baum, Alarith Uhde, Sarah Sterz, Martin C Hirsch, and Markus Langer. How Do We Assess the Trustworthiness of AI? Introducing the Trustworthiness Assessment Model (TrAM). *Computers in Human Behavior*, 170:108671, 2025. doi: 10.1016/j.chb.2025.108671.

[56] Mark Schroeder. *Reasons First*. Oxford University Press, 2021. ISBN 978-0198868224.

[57] Andre Steingrüber and Kevin Baum. Justifications For Democratizing AI Alignment And Their Prospects. In *Bridging The Gap Between AI And Reality*, volume 16220, pages 146–159, 2026. doi: 10.1007/978-3-032-07132-3_10.

[58] Zhi-Xuan Tan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in ai alignment. *Philosophical Studies*, 182(7):1813–1863, 2024. doi: 10.1007/s11098-024-02249-w.

[59] Emily Tennant, Stephen Hailes, and Mirco Musolesi. Hybrid Approaches for Moral Value Alignment in AI Agents. 2024. arXiv:2312.01818v3.

[60] Ajay Vishwanath and Marija Slavkovik. Machine Ethics or AI Alignment? In *Nais 2025: Symposium Of The Norwegian AI Society*, volume 3975 of *CEUR Workshop Proceedings*, pages 81–90, 2025.

[61] Alan F. Winfield, Katina Michael, Jeremy Pitt, and Vanessa Evers. Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems. *Proceedings of the IEEE*, 107(3): 509–517, 2019. doi: 10.1109/JPROC.2019.2900622.

[62] M.J. Wolf, K.W. Miller, and F.S. Grodzinsky. Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications. *The ORBIT Journal*, 1 (2):1–12, 2017. doi: https://doi.org/10.29297/orbit.v1i2.49.