Aggregation Problems in Machine Ethics and AI Alignment

Kevin Baum [1], **Marija Slavkovik** [2]

[1]German Research Center for Artificial Intelligence (DFKI), Germany
[2]University of Bergen, Norway
kevin.baum@dfki.de, marija.slavkovik@uib.no

Abstract

Artificial agents increasingly make decisions with far-reaching consequences. It is therefore imperative to ensure that their actions are not only functionally effective but also normatively appropriate. Two major paradigms address this challenge: machine ethics and value alignment. Machine ethics typically engages in *moral* aggregation, especially through value and (descriptive) uncertainty aggregation. Value alignment approaches tend to rely on *social* aggregation to manage value pluralism and moral uncertainty, often implicitly or indirectly. This paper disentangles these forms of aggregation and analyzes their roles across three stages of machine moral reasoning: moral evaluation, moral assessment, and moral decision. Rather than favoring one paradigm, we expose their mutual dependencies and respective blind spots, particularly under conditions of persistent moral disagreement. We argue that social aggregation cannot bypass deep normative commitments. Alignment by social aggregation cannot replace moral aggregation but merely relocates it—often opaquely.

1 Introduction

The deployment of autonomous and semi-autonomous machines raises a number of substantive challenges.[1] Beyond the technical demands of constructing capable artificial agents, there is the normative task of determining how such systems should make morally appropriate decisions—and by which standards, and according to whose judgments, their behavior should be guided.[2]

Two paradigms currently dominate attempts to meet this normative challenge: *machine ethics*, which aims to equip AI agents with explicit moral reasoning capabilities; and *value alignment*, which seeks norm-conformant behavior through implicit learning from human feedback. While these paradigms differ in methodology, both depend on some form of *aggregation* to navigate the moral landscape, including challenges of uncertainty and pluralism. Understanding how aggregation is used within each paradigm is therefore key to clarifying their respective promises and limitations.

The main paradigm in the current alignment literature is significantly shaped by what Baum (2020) has called *social-choice AI ethics*: the idea that machines should align their behavior with the moral preferences and judgments of human collectives. These preferences may be elicited directly through learning frameworks like reinforcement learning from human feedback (RLHF) (Ouyang et al. 2022), or indirectly via principles derived from fair, participatory deliberative processes (Gabriel and Keeling 2025). While these approaches are often not framed explicitly as social choice methods—and frequently overlook foundational insights from social choice theory (Conitzer et al. 2024)—they in fact instantiate (often implicit) versions of it. This trend toward social-choice-based alignment is driven, at least in part, by the persistence of moral disagreement and moral uncertainty—the absence of a commonly agreed moral ground truth. In response, social aggregation is often presented as a pragmatic way to bypass foundational ethical disputes: instead of solving moral philosophy, let machines reflect what humans want—collectively.

However, *pace* Conitzer et al. (2024) but in line with Baum (2020), we argue that social aggregation cannot avoid deep normative commitments. Even when moral theorizing is delegated to collective, social processes, key design choices remain morally significant: who has *standing* in these processes, how individual preferences and judgments are elicited and *measured*, and—most importantly—which *aggregation* procedure is chosen. In other words, aggregation is not a morally neutral technical step, but a normative decision in its own right. Shifting from explicit moral theory to human preference aggregation does not eliminate the moral terrain—it merely relocates it.

In this paper, we focus on one critical part of this terrain: the challenge of *aggregation* in moral decision-making. Our central claim is that social aggregation, while sometimes helpful, cannot replace other forms of aggregation—such as value aggregation or uncertainty aggregation—without en-

[1]Throughout this paper, we use the term "machine" to refer to any artificially intelligent or autonomous computational artifact.

[2]Typically, morality is understood broadly in these endeavors as a shared understanding of how we ought to live together, encompassing considerations of right and wrong, good and bad, virtue and vice, and our responsibilities toward one another. Besides morality, this includes questions of decency, cultural conventions, and safety considerations. Accordingly, these moral systems often lack thorough deliberation and coherence, and frequently do not withstand rigorous ethical analysis (ethics being the systematic study of morality).

countering similar normative complexities. Worse, it may obscure those complexities by hiding them inside learned models or ill-defined procedures. Our framework clarifies where ethical challenges arise in AI system design and highlights how current methods risk obscuring rather than confronting them, aiming to help researchers better locate their own contributions. Ultimately, we seek to support more principled, transparent, and accountable approaches to building morally aligned AI systems.

We proceed in four steps. First, we motivate and present a formal model of moral decision-making for artificial agents. Second, we distinguish and explicate three kinds of aggregation relevant to this setting: *value aggregation*, *uncertainty aggregation*, and *social aggregation*. Third, we locate these forms of aggregation within the structure of machine moral reasoning, showing how they arise at different stages. Finally, we explain how these types of aggregation interrelate—especially under conditions of moral uncertainty—and why this interdependence explains the allure, but also the limitations, of using social aggregation as a substitute for moral reasoning.

Related Work. The challenge of ensuring that AI agents act in (morally) permissible ways—henceforth, the *Moral Machine Challenge* (MMC)—is studied under the overlapping but methodologically distinct paradigms of machine ethics and AI alignment. This challenge is significantly complicated by the absence of a widely accepted moral ground truth and, conversely, by the persistence of moral pluralism, disagreement, and uncertainty (Robinson 2024; MacAskill, Bykvist, and Ord 2020). If there were a commonly endorsed and defensible moral theory—ideally one that was either computable or amenable to tractable approximation—it could arguably serve as a foundation for *moral aggregation*, allowing the systematic integration of morally relevant considerations into machine decision-making. However, no such theory has achieved sufficient consensus.

Although the boundaries between machine ethics and AI alignment are not always sharply defined (Baum 2025), they differ in methodological emphasis. Machine ethics focuses on embedding explicit moral reasoning mechanisms into agents (e.g., Wallach and Allen 2009; Anderson and Anderson 2011), while alignment approaches aim to ensure conformant behavior more indirectly, typically without explicit moral representations or deliberative processes (e.g., Gabriel 2020). In alignment research, the system's behavior is aligned with human preferences or values, directly via preference elicitation and reward shaping (e.g., Christiano et al. 2017; Conitzer et al. 2024), or indirectly via normative principles derived from fair deliberative procedures (e.g., Gabriel and Keeling 2025; Steingrüber and Baum 2025).

Approaches addressing the MMC as a problem of *social aggregation* have been explored across disciplines (Adler 2016; Rahwan 2018; Prasad 2018; Baum 2020; Liao et al. 2023; Ozaki, Rehman, and Slavkovik 2024; Gabriel and Keeling 2025). The core idea in these proposals is to use aggregation procedures to synthesize diverse moral preferences or judgments into actionable system behavior. These approaches aim to accommodate normative diversity without presupposing agreement on one comprehensive moral theory. However, as we argue, such aggregation strategies may shift rather than resolve foundational ethical challenges.

Contribution. In this paper, we argue that distinct forms of aggregation play essential roles in both machine ethics and value alignment approaches to the MMC. We identify and formally structure three key types of aggregation—value aggregation, uncertainty aggregation, and social aggregation—and analyze their roles within moral reasoning processes for artificial agents. Understanding how these aggregation types interrelate, and where they create normative friction or opacity, provides important conceptual foundations for addressing how machines should make decisions under conditions of moral uncertainty and disagreement.

2 Decision-Making Process for Moral AI

People make decisions both to achieve practical goals and to better understand the world (Kersten and Szpakowicz 1994). In *decision theory*, a decision problem is modeled as a set of alternatives A, typically accompanied by additional structure $f(a)$ for $a \in A$, where f is an index that evaluates or assesses each alternative with respect to its consequences, outcomes, costs, payoffs, or associated probabilities. The focus lies in modeling everything that helps to explain which alternative an individual agent should select, based on their preferences, beliefs, and objectives (Luce and Raiffa 1957). In essence, the central task is to identify an appropriate function for guiding choice.

Social choice theory extends this modeling framework to collective decision-making. Here, a decision problem includes a set of voters V, each represented by a preference relation over the alternatives in A. While preferences are typically ordinal (rankings), some models also allow for cardinal utilities. Social choice theory, thus, investigates methods for aggregating individual preferences into collective decisions, as well as the formal and conceptual limitations inherent in such aggregation.

Both decision theory and social choice theory thus focus on the *rational* selection of alternatives, albeit at different levels: the former addresses reasoning at the level of individual agents, the latter at the level of collectives.

In *moral philosophy*, decision situations are modeled with far greater variation. Some areas—especially those influenced by decision theory, and often aligned with the consequentialist tradition—analyze moral choice in terms of alternatives, consequences, and the moral evaluation of outcomes. However, much of moral philosophy focuses less on structured decision models and more on articulating and analyzing principles, rules, values, and normative reasons that determine what agents ought to do. In these traditions, the emphasis is on identifying morally salient considerations, rather than on formally structuring decision problems.

Moreover, traditional moral theory has primarily focused on individual decision-making, offering normative guidance from an isolated agent perspective.[3] In contrast, *political*

[3]There are exceptions in more specialized areas, such as col-

philosophy—especially in the tradition of public reason liberalism (e.g., Rawls 2005) and moral contractualist theory—has developed frameworks for addressing moral pluralism and disagreement. More recently, these concerns have resurfaced in the context of AI alignment, where systems must make decisions affecting multiple stakeholders with divergent moral views (Gabriel and Keeling 2025; Robinson 2024). Still, the formal modeling of genuinely collective moral decision-making remains underdeveloped, particularly with regard to the role of various forms of *aggregation*.

We argue that understanding the MMC requires more than a static specification of alternatives, preferences, and outcome appraisals. The core challenge of designing machines capable of morally appropriate decision-making—beyond mere instrumental rationality—lies not primarily in identifying *criteria of rightness*, but in developing adequate *decision procedures* (Driver 2022; Stark 1997). Addressing this challenge demands a broader view of decision-making as a dynamic, structured process that actively gathers, generates, and integrates morally relevant information.

To capture this dynamic dimension, it is helpful to draw on decision analysis and organizational psychology traditions that emphasize internal stages of decision-making. Examples include Irving Janis's GOFER model (Mann, Harmoni, and Power 2012; Janis and Mann 1977) and Kristina Guo's DECIDE framework (Guo 2008), both of which decompose decision-making into structured steps such as goal setting, option generation, information gathering, evaluation, and execution.

As in organizational and managerial contexts, the decision-making process for moral machines must be understood as systematically staged. Designing morally constrained agents requires not only specifying desirable outcomes 'somehow', but also determining where normative reasoning must enter into the structure of deliberation. Incorporating morally relevant parameters and constraints demands careful analysis of the phases in which ethical challenges arise and can be operationalized.

In the following section, we introduce a conceptual framework that identifies three key stages of moral decision-making in artificial agents, each of which requires integrating normative considerations in its own right.

3 The Structure of Moral Decisions for Machines

Designing machines capable of morally acceptable decision-making requires a refined understanding of how moral considerations are to be integrated into the decision-making process. We propose that moral decision-making for machines can be systematically analyzed by distinguishing between three nested levels: *moral evaluation*, *moral assessment*, and *moral decision*.

Let us begin by defining the concept of a decision situation. A decision situation is a state in the operation of an artificial agent in which the agent is presented with a list of options A, in a specific context c. We assume that the list of options is finite. Each option $O \in A$ can be characterized by a list of factors F. Some of these factors—let us call this subset $F_M^{O,c}$—are morally relevant in the given context c.

Moral Evaluation. Moral evaluation is the process of:

1. Identifying and considering all factors $F_M^{O,c}$, i.e., the morally relevant factors in a given decision situation (or, more precisely, to some option $O \in A$ in a given decision situation and context c).
2. Evaluating the available options with respect to the morally relevant factors that pertain to them in the specific context. This may include their consequences, their moral qualities (especially in the axiological sense of values), the principles or duties they instantiate, or other context-dependent normative considerations such as normative reasons (cf. Baum et al. 2024).
3. Associating each option with an evaluative moral status—qualitative or quantitative—that enables comparison and normative discernment.

Moral evaluation does not yet classify options according to their *deontic moral status*, i.e., as being permissible, impermissible, obligatory, or forbidden. That is the job of the next stage. Instead, moral evaluation provides the informational and normative structure required for doing so. Its goal is to make salient the morally relevant features of each option so as to enable comparison—at least in the form of a partial ranking—based on moral desirability.

Example 1. Consider a content moderation agent that is tasked with deciding whether to censor a social media post containing the photograph commonly known as the 'Napalm Girl'.[4] The agent must choose between two options: censor the post (O_1) or allow it to remain online (O_2). To determine the morally appropriate course of action, the agent must identify the morally relevant features of each option, considering both the content of the post and the intended audience of the platform it serves.

Assume that the post is intended to inform the public by curating public domain media. The photo in question raises a number of morally salient considerations. It depicts a naked child; it is a historically significant artifact; it documents the horrors of war; it belongs to the public domain; and it portrays a private individual—Phan Thi Kim Phúc—who is still alive. Each of these factors may have different moral weights depending on the broader context.

The context also contributes norms, values, and constraints. For instance: content resembling child pornography should be censored; historically significant materials should be preserved; and content should generally be left uncensored unless compelling moral reasons justify intervention.

Evaluating options O_1 and O_2 requires determining the extent to which these morally relevant factors apply to each. Censoring (O_1) may align with norms about protecting children and shielding young users from traumatic content, but it may also suppress access to historically important material and obscure the moral reality of war. Allowing the post to

lective responsibility (Smiley 2023) or the problem of collective action in consequentialism, where recent work draws on game-theoretic insights (e.g., Baum 2024).

[4] https://newyork.fotografiska.com/en/events/napalm-girl

remain (O_2) supports historical memory and informational openness, but could raise concerns about consent, dignity, and exposure to distressing imagery.

Associating a moral quality with each option could involve determining which morally relevant factors carry the greatest normative weight in the given context, or aggregating the factors into a composite evaluative judgment—whether qualitative or quantitative—that enables comparison between the options.

Moral Assessment. Based on the outcomes of moral evaluation, moral assessment involves classifying options into normative categories according to their *deontic moral status*, such as being morally permissible, impermissible, obligatory, or supererogatory—or, more practically, into context-sensitive categories such as *acceptable* or *justifiable* (cf. Baum et al. 2024). Moral assessment builds on the structured evaluation of morally relevant factors to filter the space of feasible options according to normative constraints. It involves both practical and normative reasoning, and requires a judgment-sensitive reading of the evaluative moral status of the options in light of applicable standards or principles.

Example 2. Continuing with the "Napalm Girl" example, moral assessment might determine that if historical preservation is deemed most important, O_2 (not censoring) promotes historical preservation, while O_1 (censoring) undermines it; O_2 is the only permissible option.

Alternatively, if the morally relevant factors have been aggregated into a numerical score—say, O_1 receives 3 'moral points' and O_2 receives 4—then moral assessment might apply a threshold rule, e.g., by deeming only those options with positive moral value as permissible. In that case, both options would be considered morally permissible, but O_2 would be preferred.

Moral Decision. Moral decision refers to both the act of selecting and executing an option in a morally appropriate way. Arguably, selection occurs among the options classified as permissible through moral assessment.[5] This stage may also incorporate instrumental or other considerations—such as efficiency, goal satisfaction, or morally insignificant user preferences—provided they do not override normative constraints. The agent's objectives and design criteria will typically influence how such considerations are weighed, raising meta-normative questions that are often far from trivial.

Importantly, depending on the meta-normative framework adopted, moral decision-making may also extend to situations in which no option is deemed morally permissible. Handling such moral dilemmas (McConnell 2024)—whether by selecting the least objectionable option, seeking external input, or deferring action—constitutes a morally significant process and must be treated as part of the agent's moral competence.

This decomposition of the decision-making pipeline clarifies where different aspects of moral reasoning naturally enter the moral decision-making of an artificial agent. Moral evaluation collects and organizes normative information about the situation; moral assessment applies deontic filters (Baum, Hermanns, and Speith 2019) to distinguish permissible from impermissible (or otherwise normatively classified) options; and moral decision selects and executes one of these alternatives in a norm-conforming manner.

This structure also helps to clarify the role of *moral judgments*, which are typically the outcome of moral reasoning processes and can pertain to different stages and dimensions of moral decision-making. We distinguish here between two central types: *evaluative judgments* and *deontic judgments*. Evaluative moral judgments concern what features matter morally in a given context—such as pleasure and pain, harm, preference satisfaction, fairness, autonomy, or freedom—and how these features combine to determine the moral value or quality of actions or outcomes. Deontic moral judgments, by contrast, involve classifying options according to their normative status: whether they are right, wrong, permissible, impermissible, obligatory, or supererogatory.

In practice, moral judgment thus involves at least three sub-questions: (i) which features of a decision situation are morally relevant, (ii) how these features contribute to the moral evaluation of available actions or outcomes (e.g., via principles, weighting schemes, or rules), and (iii) how such evaluations inform the deontic classification of those options. This distinction is important both conceptually and methodologically: while evaluative and deontic judgments are often intertwined, they may rely on different normative commitments and may require different forms of aggregation or justification. These judgments may be derived from a given moral theory (or sub-modules of such theories, cf. Baum (2024))—or they may be elicited through other means, such as from individual agents or groups.

Accordingly, and as we argue in more detail in the following sections, distinct types of *aggregation problems* arise at different stages of the moral decision-making process, depending on which forms of aggregation are available and appropriate. For instance, moral evaluation inherently calls for a form of *moral aggregation*—that is, the integration of multiple morally relevant considerations into an overall evaluative judgment. Similarly, later stages of the process, such as moral assessment and decision, may also require moral aggregation—e.g., to weigh competing normative criteria or to reason about the moral desirability of possible outcomes. These stages typically build on earlier aggregation, but they can also require independent aggregation functions (e.g., inter-possibility or inter-possible aggregation).

Under conditions of moral uncertainty, however, explicit moral aggregation may not be feasible or *theoretically* justifiable (Steingrüber and Baum 2025). In such cases, a common strategy in recent AI alignment work is to (propose to) replace or bypass it with *social aggregation*—for example, by synthesizing divergent human preferences, values, or judgments across relevant stakeholder groups to guide machine behavior in the absence of normative consensus.

3.1 A Formal Model of Moral Decision-Making

Before turning to aggregation-specific challenges, we introduce a formal model of moral decision-making for ma-

[5] Cf. the distinction between outcome and execution alignment discussed in (Baum 2025).

chines, grounded in the structure presented above.

Definition 1 (Moral Decision Situation (MDS) and Moral Decision Problem)**.** A *moral decision situation* (MDS) is a tuple $\langle A, c, F, f_M, a_M \rangle$, where:

- A is a finite set of available options.
- c is a context, assumed here to be a constant (for simplicity), though in practice it may consist of properties, predicates, or richer environmental information. Let C denote the space of possible contexts.
- F is a set of factors potentially relevant to the evaluation of options.
- $f_M : A \times C \to 2^F$ is a function representing *moral evaluation*, which maps each option and context to a subset $F_M^{O,c} \subseteq F$ of morally relevant factors for option $O \in A$ in context c.
- $a_M : A \times 2^F \to 2^A$ is a function representing *moral assessment*, which returns the subset $A_P \subseteq A$ of morally permissible options based on these factors.

A *moral decision problem* is then defined by a function $d_M : A_P \to 2^A$, which selects a non-empty subset $A_{\text{do}} \subseteq A_P$ for execution. In cases where $A_P = \emptyset$, d_M may fall back to a secondary mechanism (e.g., selecting the least morally objectionable option, seeking human input, or deferring action).

Moral decision situations are often conceptualized in terms of values, and thus evaluated in an absolute (cardinal) rather than a purely comparative (ordinal) manner (Dietrich and Jabarian 2022). This motivates the definition of a special case: a *value-based* MDS, which reflects this widely adopted way of modeling moral decision-making more explicitly.

Definition 2 (Value-Based MDS)**.** Let $S = \langle A, c, F, f_M, a_M \rangle$ be a moral decision situation and let $\mathbb{V}$ be a closed numerical domain representing a space of moral values.[6] S is a *value-based* MDS if the following conditions hold:

- F is a set of *value factors*, i.e., features that describe value-relevant properties of options (e.g., well-being, preference satisfaction, fairness).
- $f_M : A \times C \to 2^F$ is a *value-based evaluation function*, mapping each option and context to a subset of morally relevant factors interpreted in a narrowly axiological sense.
- a_M is a *value-based assessment function*, defined via:
  - an *option valuation function* $v_M : A \times 2^F \to \mathbb{V}^n$, which assigns to each option (given its morally relevant value factors) a vector-valued score in $\mathbb{V}^n$;[7]
  - a *value-sensitive permissibility function* $a_M^{\mathbb{V}} : A \times \mathbb{V}^n \to 2^A$, which filters morally permissible options based on the assigned values.

  That is, a_M is implemented as a two-step process: first evaluating options via v_M, then filtering them via $a_M^{\mathbb{V}}$.
- Given $\mathbb{V}$ and A_P, one may further define a (partial or total) ordering $\succ_M$ over A_P based on the values assigned by v_M.

[6] Without loss of generality, we may assume that $\mathbb{V}$ is an interval over the real numbers $\mathbb{R}$ (e.g., $[0, 1]$ or $[-1, 1]$).

[7] We allow for vector-valued evaluation to accommodate multi-dimensional moral criteria.

Example 3. Let us consider the "Napalm Girl" example again. This time, we assume that the social media company models the situation as a value-based moral decision situation in accordance with the above framework.

The set of options is $A = \{O_1, O_2\}$, where O_1 represents censoring and O_2 represents not censoring the post. The context c is a news article discussing the dangers of chemical weapons.

The factor set F may include:

f_1 hour of posting
f_2 name of author
f_3 photo of a naked child
f_4 photo of a living person
f_5 photo in the public domain
f_6 historically relevant document

The morally relevant factors are:

$$F_M^{O_1,c} = f_M(O_1, c) = \{f_3\},$$

$$F_M^{O_2,c} = f_M(O_2, c) = \{f_3, f_5, f_6\}.$$

Assume the option valuation function v_M assigns qualitative values depending on whether a factor is morally favorable ($+$), unfavorable ($-$), or neutral ($\circ$):

	f_3	f_5	f_6
O_1	$+$	$\circ$	$-$
O_2	$-$	$+$	$+$

That is, censoring (O_1) aligns with norms about protecting children (f_3), but suppresses a historically relevant document (f_6). Not censoring (O_2), by contrast, promotes public access (f_5 and f_6), but raises concerns about dignity and exposure (f_3).

An aggregated version of v_M could sum the number of morally favorable markers:

$$v_M(O_1, F_M^{O_1,c}) = 1, \quad v_M(O_2, F_M^{O_2,c}) = 2.$$

Alternatively, a priority-based function v'_M might select the value of the most normatively decisive factor. If f_6 is treated as most decisive for O_1, and f_3 for O_2, then:

$$v'_M(O_1, F_M^{O_1,c}) = -, \quad v'_M(O_2, F_M^{O_2,c}) = -.$$

In this case, both options receive the same score under v'_M, though they would be ranked differently under v_M.

Finally, assume $a_M^{\mathbb{V}}$ implements a threshold rule. If it requires at least two positive factors for permissibility, then only O_2 is deemed permissible. If it requires unanimity or gives veto power to any negative factor, then neither may be permissible.

This formal framework provides a modular structure for modeling morally competent agents. In the following sections, we examine how distinct types of aggregation problems arise within this framework, particularly at the stages of evaluation, assessment, and decision.

4 Moral Decisions under Uncertainty

There is one complication of moral decision-making that we have bracketed so far: uncertainty. Uncertainty is not a side issue—it is a central structural challenge that shapes how options are evaluated, assessed, and selected. It introduces additional degrees of freedom and, with them, design decisions that must be confronted in the development of morally competent machines.

We distinguish between two principal types of uncertainty:

Descriptive uncertainty arises from incomplete or ambiguous information about the world—for example, uncertainty about the current context, the properties of available options, the likely outcomes of actions, or the preferences and intentions of stakeholders.

Moral uncertainty arises when there is no agreement or clear guidance on which moral principles, reasons, or values should govern a decision. This includes uncertainty about what features matter morally, how they contribute to the evaluation of outcomes and options (whether under descriptive certainty or uncertainty), and how such evaluations should translate into normative assessments (e.g., whether an action is permissible or impermissible).

Both types of uncertainty can affect any stage of the decision-making process: identifying morally relevant factors, evaluating options, assessing permissibility, and selecting an action. Crucially, they motivate the need for *uncertainty aggregation*—procedures that synthesize evaluations across multiple possible contexts, outcomes, or normative perspectives.

To illustrate how these types of uncertainty apply in the context of the MMC, we return to our earlier example.

Example 4. Returning to our example of a content moderation system evaluating whether to censor a post containing the "Napalm Girl" photograph, we can now illustrate how the different types of uncertainty discussed above map onto distinct stages of the moral decision-making process, as formalized in our MDS framework.

Let the available options be $A = \{O_1, O_2\}$, where O_1 represents censoring the post and O_2 represents allowing it. Let $S = \langle A, c, F, f_M, a_M \rangle$ be a (value-based) moral decision situation for this case, extended with valuation, assessment, and decision functions v_M, $a_M^{\mathbb{V}}$, and d_M.

While descriptive uncertainty can be reduced to uncertainty about the actual context of the decision situation, it is helpful to distinguish several distinct points at which *moral uncertainty* enters the picture:

Descriptive uncertainty affects f_M. The system may (and in practice typically will) lack full information about the context, such as the post's audience, intent, or geopolitical framing. Since the correct context $c \in C$ is unknown, f_M must consider several plausible contexts $c_1, \ldots, c_k$. Which context obtains influences which morally relevant factors $F_M \subseteq F$ are identified—whether these are (momentary or historical) factual properties (e.g., the historical significance of the photo, or whether removing it would actually constitute censorship in a (legally or) morally relevant sense), or forward-looking properties (e.g., expected harm caused by allowing it). f_M must be designed to integrate multiple plausible contexts, ideally aggregating them in a way that produces an epistemically robust set of morally relevant factors.

Moral valuation uncertainty arises both in f_M and v_M. Even if the context were known, different moral theories may prioritize different morally relevant factors or evaluate the same factors differently. For example, some theories may emphasize harm minimization or dignity, while others prioritize historical memory or freedom of expression. Rather than selecting a single normative perspective, the system could integrate across multiple. Either way, these functions must aggregate the results into a unified value assignment for each option.

Moral assessment uncertainty concerns the permissibility function $a_M^{\mathbb{V}}$. Even if scalar values have been assigned to each option, the criteria for translating these values into deontic classifications (e.g., permissible vs. impermissible) may themselves be unclear or contested. Thresholds or comparative principles may be underdetermined or disputed. Again, rather than fixing one such mapping, $a_M^{\mathbb{V}}$ could aggregate across plausible evaluative-to-deontic interpretations.

Moral decision uncertainty affects the function d_M, which selects from the set of morally permissible options $A_P \subseteq A$. If both O_1 and O_2 are permissible under the prior stages, d_M still requires a decision. This may reflect further normative priorities or stakeholder preferences—for instance, a liberal view might prioritize the individual preferences of the decision-maker (e.g., platform policy), while another might emphasize optimizing for user trust or public accountability. Rather than embedding one such stance directly into d_M, the function could implement a choice aggregation procedure to synthesize across multiple, potentially competing priorities.

This example illustrates how our MDS framework can, in principle, accommodate different types of uncertainty at different stages of moral reasoning, without requiring arbitrary selection between plausible alternatives. At this point, we leave open how such integration is to be formally modeled, as the next section turns to a more detailed discussion of the various kinds of aggregation problems.

5 Aggregation Problems in Moral Machine Decision-Making

In the previous section, we outlined the structure of moral decision-making for machines, distinguishing between moral evaluation, moral assessment, and moral decision. Each of these stages introduces distinct challenges regarding the integration of normative considerations into the agent's deliberative process. Crucially, many of these challenges intuitively involve *aggregation*: the task of combining values, information, judgments, or inputs that may be diverse, incomplete, or conflicting.

5.1 Aggregation Functions

An aggregation function, in mathematics, is a function that takes a set or vector of values and returns a single summary value. Typical examples include *max*, *min*, *sum*, or *average*. In social choice theory, aggregation functions are understood more broadly as procedures for deriving collective choices from individual opinions, preferences, or judgments—for example, selecting the option supported by the largest number of individuals.

To distinguish between these interpretations, we define three types of aggregation functions relevant to our framework: *value aggregation functions*, which combine multiple morally relevant evaluations into a single value; *choice aggregation functions*, which derive socially endorsed options from a profile of individual preference orderings;[8] and *uncertainty aggregation functions*, which handle both descriptive and moral uncertainty.

We begin with value aggregation:[9]

Definition 3 (Value Aggregation Function)**.** Let $\mathbb{V}$ be a numerical domain of moral values. A *value aggregation function* is a function

$$v : \bigcup_{n \in \mathbb{N}} \mathbb{V}^n \to \mathbb{V}$$

that assigns to each finite sequence of morally relevant values an overall value, and satisfies at least the following conditions:

Idempotence: For all $x \in \mathbb{V}$: $v(x) = x$.

Neutrality: If $y \in \mathbb{V}$ is morally neutral, then for any sequence $(x_1, \ldots, x_k) \in \mathbb{V}^k$,

$$v(x_1, \ldots, x_k) = v(x_1, \ldots, x_k, y).$$

Positive responsiveness (monotonicity): If $y \in \mathbb{V}$ is morally positive, then

$$v(x_1, \ldots, x_k) < v(x_1, \ldots, x_k, y).$$

Pareto: If two sequences differ only at position i, and $x_i \geq x_i'$, then

$$\begin{aligned} &v(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) \\ \geq\ &v(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \end{aligned}$$

While some have challenged whether value aggregation is necessary or even appropriate for moral decision-making (most famously Taurek (1977)), each of the properties in the above definition captures a minimal and widely accepted normative constraint on how morally relevant values should be combined—if they are to be combined at all. *Idempotence* ensures that if there is only a single morally relevant value, the aggregator returns that value unchanged. This anchors the function in the simple base case and avoids distortion in degenerate inputs. *Neutrality* reflects the idea that some values may be morally inert—neither improving nor worsening an outcome. Including such a neutral value should not affect the result. *Positive responsiveness (monotonicity)* ensures that adding a morally positive value must increase the overall evaluation, thereby enforcing the minimal moral significance of improvements. *Pareto* requires that if nothing is worse in one situation than in another, then the first situation cannot be worse than the second.

There are other plausible candidates for properties of value aggregation functions—such as *anonymity*, which demands that the ordering of input values does not affect the result[10]—which may also be normatively attractive, particularly in interpersonal or symmetric moral settings. However, we do not include them here as minimal conditions, since their normative force may depend on how the input values are interpreted.

Second, we may need to aggregate scalar values of *different types*, such as those reflecting privacy, freedom, welfare, or fairness. This is a case of *inter-value* aggregation, which arises in settings characterized by value pluralism. The challenge here is to combine heterogeneous values that are morally significant in fundamentally different ways. For an illustration of this second kind, we return to the "Napalm Girl" case:

Example 5. We can reinterpret the morally relevant factors in terms of distinct value types. For instance, f_3 (photo of a naked child) may correspond to concerns of dignity and privacy; f_5 (public domain) to legal permissibility or access rights; and f_6 (historical relevance) to values such as truth-telling or the preservation of collective memory.

When aggregated to determine the moral value of an option, these represent a case of *inter-value aggregation*: combining values of fundamentally different kinds. The challenge here is not merely how to sum up the interests or welfare of different individuals—as in interpersonal aggregation—but how to make trade-offs between (potentially) incommensurable values such as dignity versus truth, or privacy versus informational integrity.

Different implementations of the option valuation function v_M reflect different approaches to this aggregation problem. One might encode all values in a vector format that allows prioritization via a value-sensitive permissibility function $a_M^{\mathbb{V}}$ using, for instance, lexicographic ordering. Another might incorporate conditional, context-dependent hierarchies (e.g., freedom of expression outweighs privacy only in cases of overriding public interest). All of these are instances of value aggregation—but not over a homogeneous value type.

Importantly, both kinds of aggregation—interpersonal and inter-value—fall under the umbrella of value aggregation as introduced here. In real-world moral decision-making, they may also arise together, requiring integrated aggregation strategies that are sensitive to both individual distribution and value pluralism.

While value aggregation focuses on integrating morally relevant considerations from a single evaluative perspective, many real-world applications of moral machine reasoning

[8] For basic formal desiderata governing such functions, we follow the treatments in (Grabisch et al. 2009) and (Arrow 2012).

[9] Note that the domain $\mathbb{V}$ may contain scalar- or vector-valued moral representations, depending on the modeling context.

[10] **Anonymity:** For any permutation π of the indices $\{1, \ldots, n\}$, $v(x_1, \ldots, x_n) = v(x_{\pi(1)}, \ldots, x_{\pi(n)})$.

involve collective input: multiple agents, stakeholders, or voters may hold different—and possibly conflicting—views about which actions are morally preferable. In such settings, we require a different kind of aggregation: *choice aggregation*, which synthesizes a set of individual preference orderings into a collective decision. For this, we adopt a standard social choice framework in which individual preferences are modeled as partial orders over a shared set of alternatives (a *profile*), and collective choices are derived via an aggregation function.

Definition 4 (Choice Aggregation Function)**.** Let A be a finite set of options (also called *alternatives*). Let $A_{\succ}$ denote the set of all partial orders over A (i.e., binary relations that are reflexive, antisymmetric, and transitive).

A *choice aggregation function* is a function

$$f : A_{\succ}^n \to 2^A,$$

which takes as input a *profile* $P = (\succ_1, \ldots, \succ_n)$ of n individual partial orderings and returns a non-empty subset $f(P) \subseteq A$ as the collective choice.

Unlike value aggregation functions, there is no canonical characterization of choice aggregation functions. Since the goal is to identify the most representative option(s), a variety of normative conditions have been proposed. One particularly defensible condition is:

Unanimity: If there exists $a \in A$ such that $a \succ_i a'$ for all $a' \in A \setminus \{a\}$ and for all $i = 1, \ldots, n$, then $f(P) = \{a\}$.

Note that the notion of unanimity used here plays a role analogous to idempotence in the context of value aggregation, but the two conditions differ in form and interpretation. This formulation follows standard conventions in the social choice literature.

It is common to adopt further normative properties such as Pareto optimality, independence of irrelevant alternatives, anonymity, and others. However, not all desirable properties can be satisfied simultaneously, and the choice of which conditions should or can be imposed is closely guided by the nature and context of the decision the function is meant to support. We return to a discussion of these trade-offs in Section 6.

Finally, we turn to the challenge of aggregating uncertainty. As discussed in the previous section, moral evaluations and assessments frequently depend on both descriptive and moral uncertainty. At least some such cases call for a form of aggregation not captured by the two kinds introduced so far: value aggregation over morally relevant features, and choice aggregation over preference profiles. Instead, what is needed is aggregation across a space of possible contexts (capturing descriptive uncertainty) and evaluative perspectives (capturing normative uncertainty).

One principled approach to this challenge draws on the structure of expected utility theory: given a set of relevant contexts and value functions, each weighted by a degree of credence, uncertainty aggregation aims to combine them into a single meta-evaluation. The following definition captures this idea:[11]

Definition 5 (Uncertainty Aggregation Function)**.** Let $\mathbb{V}$ be a numerical domain of moral values. Let A be a finite set of options (also called *alternatives*), and let C be a set of contexts. Let $V = \{v_1, \ldots, v_n\}$ be a finite set of valuation functions $v_i : A \times C \to \mathbb{V}^n$, representing distinct moral theories or evaluative stances. Let

$$\Pr : C \times V \to [0, 1]$$

be a joint probability distribution over contexts and evaluative perspectives.

An *uncertainty aggregation function* is a function

$$u_{\Pr} : A \to \mathbb{V}^n$$

defined as:

$$u_{\Pr}(a) = \sum_{(c,v) \in C \times V} \Pr(c, v) \cdot v(a, c)$$

for each option $a \in A$. This yields a *meta-value* for each option, reflecting its expected moral value under both descriptive and normative uncertainty.

While other—potentially non-aggregative—approaches to integrating descriptive and normative uncertainty (and risk) into moral decision-making may exist, uncertainty aggregation functions of the kind defined here arguably play a central role in value-based moral decision-making under uncertainty, as they enable the systematic combination of evaluations across possible contexts (and their associated outcomes) as well as across competing normative assumptions and evaluative perspectives. This raises the question of how the different types of aggregation introduced so far—value, choice, and uncertainty—fit into the structured model of moral decision-making we developed earlier. We now return to the formal MDS framework to clarify their respective roles.

5.2 Aggregation and Moral Decision Situations

We now revisit our formal model of a moral decision situation (MDS) and argue that several of its constituent functions—or natural extensions thereof—either instantiate or rely on distinct forms of aggregation.

f_M: While f_M is formally defined as a mapping from options and contexts to morally relevant factors, it may require aggregation in at least two kinds of cases involving uncertainty. First, under moral uncertainty, where multiple moral perspectives yield different criteria of relevance, f_M must be supplemented by a mechanism that synthesizes these perspectives into a coherent factor set. Second, under descriptive uncertainty, when the context itself is ambiguous or partially observed, f_M may depend on a process that integrates across several plausible contextual states. In both cases, this introduces a form of *uncertainty aggregation*, producing a moral input suitable for downstream evaluative or deontic reasoning.

[11] A comprehensive taxonomy of uncertainty aggregation under the label "Expectationalism", which also distinguishes between ex-post and ex-ante approaches and allows for modeling dependencies between descriptive and normative uncertainty, can be found in (Dietrich and Jabarian 2022).

v_M**:** The option valuation function v_M, which assigns a moral value to an option based on its morally relevant features, is a paradigmatic case of *value aggregation*. It combines multiple morally significant factors—such as well-being (potentially across individuals), fairness, or autonomy—into a single evaluative score that supports comparison and ranking.[12]

$a_M^{\mathbb{V}}$**:** The value-sensitive permissibility function $a_M^{\mathbb{V}}$ determines which options are deemed morally permissible based on their aggregated moral values. In cases where the mapping from value to permissibility is itself uncertain—for instance, due to indeterminate thresholds or contested evaluative principles—$a_M^{\mathbb{V}}$ may depend on a form of *uncertainty aggregation*, synthesizing judgments across competing normative interpretations. These interpretations might, in turn, be reconciled through a higher-order application of *choice aggregation*.

d_M**:** The moral decision function d_M selects one or more options from the set of permissible ones. In contexts involving multiple stakeholders, conflicting goals, or plural evaluative perspectives, this function may require a form of *choice aggregation*, for instance, to prioritize among permissible options based on stakeholder input or external criteria.

This analysis highlights that different stages of moral decision-making necessarily or contingently involve distinct types of aggregation, depending on the kind of input being combined and the form of uncertainty or pluralism at stake. Recognizing whether aggregation is evaluative, social, or uncertainty-driven is essential for clarifying what kinds of justifiable design decisions and normative commitments each stage demands. In the following sections, we explore these types of aggregation in more detail, beginning with moral aggregation and its challenges.

5.3 Moral Aggregation and Social Aggregation

Aggregation problems in moral machine decision-making arise at two structurally distinct levels: within the agent's own moral reasoning, and in its interaction with plural human inputs. We refer to these as *moral aggregation* and *social aggregation*, respectively.

The distinction is reflected in the literature. For instance, Hirose (2014) defines moral aggregation as a "trade-off between benefits to a group of individuals and losses to another group," thus highlighting the challenge of integrating morally significant factors into a single evaluation. In contrast, social choice theory—as formalized in, e.g., Brandt et al. (2016)—addresses how to derive collective decisions from a profile of individual judgments or preferences.

We propose the following working definitions:

- **Moral Aggregation**: Within moral evaluation, the agent must aggregate different morally relevant factors—such as the consequences of actions, duties, rights, claims, reasons, or values—to form an integrated evaluation of the decision situation. This type of aggregation operates over features of the options themselves and their morally salient aspects. Strictly speaking, the aggregation must deliver a sufficiently rich structure to enable moral assessment, which ultimately boils down to determining the set of permissible options. In particular, this does not necessarily require aggregating the morally relevant factors into a single real number (as sometimes assumed, cf. (Hirose 2014)), or even into any quantitative structure at all (such as vectors, cf. (Mill 1861; Sen 1980)).[13]
- **Social Aggregation**: Under conditions of moral uncertainty and disagreement, especially when relying on human input, the agent must aggregate divergent judgments or preferences provided by different stakeholders, whether individuals or groups. This aggregation is social in nature: it concerns the integration of plural human perspectives into the machine's normative reasoning. It can be learned implicitly (Christiano et al. 2017), implemented through functions inspired by social choice theory (Conitzer et al. 2024), or guided by discourse-oriented procedural approaches, traditionally discussed in political philosophy (Rawls 1985, 1995; Habermas 1990) and more recently in AI alignment (Gabriel 2020; Gabriel and Keeling 2025).

While moral aggregation concerns the integration of morally relevant information, social aggregation becomes necessary when the machine must reason under pluralism and normative uncertainty. Both forms of aggregation are essential for constructing morally acceptable machine behavior, but each introduces specific difficulties. Moral aggregation must address how diverse normative factors—assuming they can be identified—should be weighted, compared, or combined; social aggregation must address how human disagreement and divergence of judgment should be processed.

In the following sections, we take a closer look at the normative structure of aggregation in moral machine decision-making. We examine key constraints that govern both moral and social forms of aggregation—highlighting when aggregation can justifiably be used, and when it risks overriding morally significant considerations. Our analysis shows that aggregation, while indispensable, must be treated as a morally loaded process. Designing AI agents capable of justifiable, contestable, and corrigible moral decision-making thus requires understanding not just how to aggregate, but when—and on what terms—it is appropriate to do so.

6 Normative Constraints on Aggregation

Once we accept that aggregation plays an essential role in moral decision-making for machines, the next question is how aggregation should proceed. This question arises both for value aggregation—how morally relevant factors are combined into evaluations—and for choice aggregation—how options are selected based on moral or stakeholder inputs. In both cases, normative properties of the aggregation function shape what kinds of decisions can be justified.

In the social choice tradition, aggregation has been extensively studied through a formal lens. Foundational work by

[12]The moral value may be scalar or vector-valued depending on the domain $\mathbb{V}$, as discussed above.

[13]For a non-aggregative extension of maximizing act-consequentialism as an example, see (Baum 2024).

Arrow (2012) identifies a set of desirable properties for preference aggregation functions: independence of irrelevant alternatives (IIA), the Pareto principle, and non-dictatorship.[14] Arrow famously showed that no aggregation function can satisfy all three simultaneously, which underscores the importance of identifying context-specific normative priorities.

Conitzer et al. (2024) revisits this problem for AI ethics, emphasizing properties like *independence of clones* and *strategy-proofness* for moral decision-making. They also highlight *anonymity*—requiring that participant identity not affect outcomes. While properties like strategy-proofness and clone independence are relatively uncontroversial, anonymity raises particular challenges in moral contexts. To see why, consider cases where identity carries moral significance: we might give a birthday person priority in lunch ordering, or recognize how cultural background and lived experience shape moral perspectives. This raises questions about when individual inputs deserve special priority.

This brings us to a normative constraint less frequently formalized, but no less important: the *liberalism condition*, introduced by Sen (1970). In its simplest form, liberalism requires that some aspects of a decision be left to the authority of individuals rather than the aggregation function. Sen formalized this as the idea that certain individuals should have decisive power over particular issues. Although his famous "liberal paradox" shows that this condition cannot be universally guaranteed alongside other desirable properties, it remains a powerful normative intuition—especially in moral decision-making.

Its relevance for AI becomes clear when we consider machines designed to act as personal moral agents. Think of a wearable device, smartphone, or household robot. These systems may be designed to reflect not collective values, but the moral perspective of a specific user. For example, someone who considers nudity intrinsically wrong may want their smartphone to flag or block such images. Aggregating across many users to define what counts as 'morally acceptable' in this case would violate their autonomy. In such cases, liberalism means allowing individuals to constrain or override the system's behavior based on their own evaluative commitments.

Even in shared or public systems—like self-driving cars, public-facing recommender systems, or medical triage tools—there may be principled reasons to grant certain agents standing to block or shape decisions. Liberalism in these contexts may not follow from individual control, but from institutional roles (e.g., a doctor's judgment) or democratic legitimacy (e.g., a stakeholder representative's veto).

Taken together, these observations clarify a key result of our analysis: moral decision-making for machines cannot rely on aggregation alone. While aggregation is necessary—to integrate diverse moral considerations, to reconcile pluralism, and to handle uncertainty in a principled manner—it must be guided by substantive normative judgments *about* aggregation, e.g., which values deserve priority, which voices deserve standing, and which decisions are properly left to individual discretion. These choices are not merely technical but *deeply* moral. In the final section, we return to the broader challenge: can social aggregation ever replace moral evaluation—or must we treat both as indispensable to morally competent artificial agents?

[14] We omit technical definitions here, since they would distract from the conceptual point.

7 Summary

One might hope to bypass moral and uncertainty aggregation challenges through social aggregation—procedures synthesizing judgments and preferences of individuals or groups. This approach, common in AI alignment work (Gabriel 2020; Gabriel and Keeling 2025), suggests delegating moral evaluation to public and fair, collective processes when agreement on moral features or evaluation methods is unavailable.

However, as Baum (2020) argues, this strategy merely relocates the site of moral decision-making. Social aggregation itself rests, as many other morally significant design decisions, on a series of *positive normative choices* (Wachter, Mittelstadt, and Russell 2021): Who counts as a stakeholder? How are inputs represented and weighed? Which aggregation procedure with which properties is used? These design choices are not morally neutral—they structure the outcome, and must themselves be normatively justified.

The implication is not that social aggregation should be rejected, but that it must be treated as a *morally charged* process in its own right. In the absence of consensus on substantive moral principles, the best we can aim for is a form of decision-making that makes its assumptions explicit, allows for scrutiny, and remains open to revision.

In this paper, we have proposed a structured view of aggregation in moral machine decision-making, introducing three types—value aggregation, choice aggregation, and uncertainty aggregation—and mapping them to our formal moral decision situation framework. We explored normative constraints on these functions and examined risks in deferring moral reasoning to social aggregation.

The upshot is that aggregation is not a substitute for moral reasoning, but a morally embedded design problem. Building morally competent machines requires identifying appropriate aggregation types for each stage and context. Such systems must embody transparency, contestability, and corrigibility: their decisions must be traceable to explicit normative justifications, open to challenge, and responsive to new perspectives.

Acknowledgments

The work of Kevin Baum is partially funded by DFG grant 389792660 as part of TRR 248 – CPEC, see https://perspicuous-computing.science, by the German Federal Ministry of Education and Research (BMBF) as part of the project *MAC-MERLin* (Grant Agreement No. 01IW24007), and the European Regional Development Fund (ERDF) as well as the German Federal State of Saarland within the scope of the Project *ToCERTAIN*. The work of Marija Slavkovik is partially funded by Trond Mohn forskningsstiftelse (grant no. TMS2023TMT01).

References

Adler, M. D. 2016. Aggregating Moral Preferences. *Economics and Philosophy*, 32(2): 283–321.

Anderson, M.; and Anderson, S. L. 2011. *Machine Ethics*. Cambridge University Press.

Arrow, K. J. 2012. *Social Choice and Individual Values*. Yale University Press. ISBN 9780300179316.

Baum, K. 2024. *Doing Wrong with Others: Multi-Agent Consequentialism as a Solution for the Collective Action Problem*. Ph.D. thesis, Universitätsbibliothek Dortmund.

Baum, K. 2025. Disentangling AI Alignment: A Structured Taxonomy Beyond Safety and Ethics. In *Bridging the Gap Between AI and Reality: First International Conference, AISoLA 2024, Crete, Greece, October 30 – November 03, 2024, Selected Papers*. Berlin, Heidelberg: Springer.

Baum, K.; Dargasz, L.; Jahn, F.; Gros, T. P.; and Wolf, V. 2024. Acting for the Right Reasons: Creating Reason-Sensitive Artificial Moral Agents. In *Workshop on Formal Ethical Agents and Robots (FEAR)*.

Baum, K.; Hermanns, H.; and Speith, T. 2019. Towards a framework combining machine ethics and machine explainability. *arXiv preprint arXiv:1901.00590*.

Baum, S. 2020. Social Choice Ethics in Artificial Intelligence. *AI & Society*, 165–176.

Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D. 2016. *Handbook of Computational Social Choice*. USA: Cambridge University Press, 1st edition. ISBN 1107060435.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Conitzer, V.; Freedman, R.; Heitzig, J.; Holliday, W. H.; Jacobs, B. M.; Lambert, N.; Mossé, M.; Pacuit, E.; Russell, S.; Schoelkopf, H.; et al. 2024. Position: social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, 9346–9360.

Dietrich, F.; and Jabarian, B. 2022. Decision under normative uncertainty. *Economics & Philosophy*, 38(3): 372–394.

Driver, J. 2022. Moral Theory. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.

Gabriel, I.; and Keeling, G. 2025. A matter of principle? AI alignment as the fair treatment of claims. *Philosophical Studies*.

Grabisch, M.; Marichal, J.-L.; Mesiar, R.; and Pap, E. 2009. *Aggregation Functions (Encyclopedia of Mathematics and its Applications)*. USA: Cambridge University Press, 1st edition. ISBN 0521519268.

Guo, K. L. 2008. DECIDE: A Decision-Making Model for More Effective Decision Making by Health Care Managers. *The Health Care Manager*, 27(2): 118–127.

Habermas, J. 1990. *Moral Consciousness and Communicative Action*. MIT press.

Hirose, I. 2014. *Moral Aggregation*. New York, US: Oxford University Press US.

Janis, I. L.; and Mann, L. 1977. *Decision making: A psychological analysis of conflict, choice, and commitment.* Free press.

Kersten, G. E.; and Szpakowicz, S. 1994. Decision making and decision aiding: Defining the process, it representations, and support. *Group Decision and Negotiation*, 3(2): 237–261.

Liao, B.; Pardo, P.; Slavkovik, M.; and van der Torre, L. 2023. The Jiminy Advisor: Moral Agreements among Stakeholders Based on Norms and Argumentation. *J. Artif. Intell. Res.*, 77: 737–792.

Luce, D. R.; and Raiffa, H. 1957. *Games and Decisions*.

MacAskill, M.; Bykvist, K.; and Ord, T. 2020. *Moral Uncertainty*. Oxford University Press.

Mann, L.; Harmoni, R.; and Power, C. 2012. The GOFER course in decision making. In *Teaching decision making to adolescents*, 61–78. Routledge.

McConnell, T. 2024. Moral Dilemmas. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2024 edition.

Mill, J. S. 1861. *Utilitarianism*. Cleveland: Oxford University Press UK.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Ozaki, A.; Rehman, A.; and Slavkovik, M. 2024. Finding middle grounds for incoherent horn expressions: the moral machine case. *Autonomous Agents and Multi-Agent Systems*, 38(2): 50.

Prasad, M. 2018. Social Choice and the Value Alignment Problem. In *Artificial Intelligence Safety and Security, ed. R. V. Yampolskiy*, 291–314. Boca Raton, FL: Chapman and Hall/CRC. ISBN 9781351251389.

Rahwan, I. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1): 5–14.

Rawls, J. 1985. Justice as Fairness: Political Not Metaphysical. *Philosophy and Public Affairs*, 14(3): 223–251.

Rawls, J. 1995. Political Liberalism: Reply to Habermas. *The Journal of Philosophy*, 92(3): 132–180.

Rawls, J. 2005. *Political Liberalism: Expanded Edition*. Columbia University Press.

Robinson, P. 2024. Moral disagreement and artificial intelligence. *AI & Society*, 39(5): 2425–2438.

Sen, A. 1970. The Impossibility of a Paretian Liberal. *Journal of Political Economy*, 78(1): 152–157.

Sen, A. 1980. Plural Utility. *Proceedings of the Aristotelian Society*, 81: 193–215.

Smiley, M. 2023. Collective Responsibility. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition.

Stark, C. 1997. Decision Procedures, Standards of Rightness and Impartiality. *Nous*, 31(4): 478–495.

Steingrüber, A.; and Baum, K. 2025. Justifications for Democratizing AI Alignment and Their Prospects. In *Bridging the Gap Between AI and Reality: Third International Conference, AISoLA 2025, Proceedings*. Berlin, Heidelberg: Springer. In print, arXiv:2507.19548v1.

Taurek, J. M. 1977. Should the numbers count? *Philosophy & Public Affairs*, 293–316.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *West Virginia Law Review*, 123(3).

Wallach, W.; and Allen, C. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press. ISBN 9780195374049.