

Values & Preferences in Reasoning

*2026 Symposium Pre-Proceedings. Hosted at the Artificial
Intelligence and Simulation of Behaviour Convention 2026.*

Editors:

Jay Paul Morgan

Adam Wyner

Contents

1	Argument over combinations of actions	4
	<i>Frederico Castagna, Katie Atkinson, and Peter McBurney</i>	
2	Formalizing Value Ontologies for Smart-City Governance in Isabelle/HOL: Fixed and Context-Sensitive Preference Reasoning with Ceteris Paribus	14
	<i>Andrea Vestrucchi, Philip Wölde</i>	
3	From Ethical Breakdown to Normative Safety: Five Failure Models for Responsible Automated Decision-Making	26
	<i>Ermo Tåks</i>	

Argument over combinations of actions

Federico Castagna^{1,*}, Katie Atkinson² and Peter McBurney³

¹*Brunel University, Uxbridge, London UB8 3PH UK*

²*University of Liverpool, Brownlow Hill, Liverpool L69 7ZX UK*

³*King's College London, 30 Aldwych, London WC2R 2LS UK*

Abstract

This paper examines how argumentation-based justifications for actions can be extended from single actions to combinations of actions. Prior work presented a practical-reasoning based argumentation scheme for a proposed action in terms of the goals that executing the action would achieve and the social values which these goals then promoted or demoted. We build on this approach to investigate how such justifications can be systematically composed when actions are combined, drawing on a typology of composition of actions from Propositional Dynamic Logic such as sequence, iteration, and parallel execution. For each type of combination, we identify the logical constraints required to ensure that individual actions may be feasibly combined and we show how goals, states, and value effects interact in the resulting combined justifications. This work provides a step toward richer models of practical reasoning, supporting more complex forms of deliberation and coordination in multi-agent systems.

Keywords

computational argumentation, practical reasoning, argumentation schemes, multi-agent systems

1. Introduction

In everyday conversations, arguments are generally used to advocate for or assert a conclusion based on premises presented as evidence. These arguments do not always conform to the formal structures typically associated with deductive or inductive reasoning. Instead, they are often referred to as “plausible” or “presumptive” indicating that they are only tentatively or provisionally accepted, even when they are accurate [1]. Argument schemes are inference structures that enable the identification and evaluation of common types of presumptive argumentation in everyday discourse [2]. Argument schemes are usually paired with a set of specific critical questions that provide the ability to probe and assess the given argument arising from instantiation of the scheme in a particular case, in relation to the context in which the argument has been put forward¹.

Intelligent agents must engage in *practical reasoning* [4] to effectively interact with their environment and determine the most appropriate course of action in any given situation. Therefore, an inference pattern that formalises the rationale behind this process could be particularly valuable in delivering explainable AI and automated reasoning. Argument schemes and critical questions have been used to computationally model a variety of types of reasoning [5]. In terms of practical reasoning, they have been used to provide a justification for a proposed course of action on the basis of the social values that executing the action would promote or demote. For instance, in [6], the authors present the following argument scheme to justify an action α :

Executing action α in State R takes us to State S , where a set of propositions (Goals) G is true, and which, becoming true, thereby promote Value v .

Values & Preferences in Reasoning Symposium (VPR), 2026, July 01–02, 2026, University of Sussex, Brighton, UK

*Corresponding author.

[†]These authors contributed equally.

✉ federico.castagna@brunel.ac.uk (F. Castagna); K.M.Atkinson@liverpool.ac.uk (K. Atkinson); peter.mcburney@kcl.ac.uk (P. McBurney)

ORCID 0000-0002-5142-4386 (F. Castagna); 0000-0002-5683-4106 (K. Atkinson); 0000-0001-5547-5239 (P. McBurney)

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

¹A different strand of research argues for disentangling argument schemes and critical questions, as it is possible to have the latter without the former [3].

Goals represent the effects that an agent wishes to attain, whereas *States* are the consequences of the undertaken action, whether the agent desires them or not. *Values*, in turn, provide the actual reasons for which an agent wishes to achieve a goal [6]. In symbols, this justification can be presented as:

$$\alpha : R \rightarrow S \models G \uparrow v$$

Consequently, the proposed action α can be justified or attacked on the basis of such a scheme. This representation also allows for justifications which demote (rather than promote) a value, or in symbolic form:

$$\alpha : R \rightarrow S \models G \downarrow w$$

A proposed action could therefore be justified on the basis of the application of such a scheme. The critical questions associated with this scheme provide the basis for reasoned attack (and subsequent defence) on the proposed action yielded from instantiation of the scheme in a given context.

Since [6], semantics for this approach have been given in terms of Alternating-Time Temporal Epistemic Logic (ATEL) in [7] and in terms of category theoretic structures in [8]. Applications of the approach have also been developed, for example, for analysis of discussions between software developers about software design proposals [9] or for machine-to-machine dialogs involving commands [10]. The legal domain has also served as a key area in which the scheme has been shown to be usefully deployed [11].

Furthermore, prior work in [12] has investigated how argumentation-based dialogues can be used as a way to coordinate actions in multi-agent planning scenarios. Building on this line of inquiry, this paper considers how argument-scheme based justifications for actions may be *integrated* when the actions are proposed to be combined. The overall aim of the paper is to extend the practical-reasoning argumentation scheme justification for an action as set out in [6] to generate a similar justification for the combination of two or more actions.

2. Combining Actions

How may actions be combined? As a reference point, we consider the possible ways in which two actions may be combined in Propositional Dynamic Logic (PDL), a modal propositional logic developed to reason about computer programs and their composition [13]. In PDL, actions may be combined as follows:

- **Iteration:** Action α is executed sequentially a finite number of times. We symbolize this combination by repeated naming of the action, separated by commas, thus: α, α, α .
- **Sequence:** Action α is executed and then action β is executed. We symbolize this combination by naming the two actions in order of execution, separated by a comma, thus: α, β .
- **Parallel:** Two actions α and β are executed starting at the same time. We symbolize this combination by naming the two actions separated by an ampersand, thus: $\alpha \& \beta$.
- **Choice:** Either action α is executed or action β is executed (but not both). We symbolize this combination by naming the two actions separated by an "exclusive or" symbol, thus: $\alpha \vee \beta$.
- **Interleaving:** Some part of action α is executed followed by some part of action β followed by some further part of action α , and so on. We symbolize this combination by naming the two actions separated by parallel vertical lines, thus: $\alpha \parallel \beta$.

3. Combining arguments for actions

We now consider the logical constraints on the combinations of arguments for proposals for actions when we wish to combine actions in the various ways of combination of the PDL typology. In each

case, we then articulate an argument-scheme based justification for the particular combination of actions building on the justifications of each of the constituent actions in the combination.

3.1. Combining arguments for sequences of actions

Suppose we wish to provide a justification for a combination of two actions, where we are given separate justifications for each action. Does the combination of actions generate any logical constraints on the combination of justifications?

For example, assume we are given the following justifications for two actions α and β (which we will assume are two different actions):

$$\begin{aligned}\alpha : R \rightarrow S &\models G \uparrow v \\ \beta : T \rightarrow U &\models H \uparrow w.\end{aligned}$$

Then, in order to execute action α and β in sequence, we must have:

State S contains state T . In other words, the set of propositions represented by state T must be a subset of the set of propositions represented by state S .

If this condition holds, then we may combine the argument schemes for the sequence of α followed by β as follows:

$$\alpha, \beta : R \rightarrow S \supseteq T \rightarrow U \models (G \cup H) \uparrow v \uparrow w.$$

Alternatively, suppose that for action β we have an argument that this action leads to the demotion of value w , i.e., $\beta : T \rightarrow U \models H \downarrow w$. Then, the combined justification for the sequence of α followed by β would be:

$$\alpha, \beta : R \rightarrow S \supseteq T \rightarrow U \models (G \cup H) \uparrow v \downarrow w.$$

If action α promotes value v while action β demotes the same value v , we can present the following argument, even if we are not able to determine which of the two consequences, promotion or demotion of v , has the stronger effect.

$$\alpha, \beta : R \rightarrow S \supseteq T \rightarrow U \models (G \cup H) \uparrow v \downarrow v.$$

3.2. Combining arguments for iterations of actions

The case of an iteration of an action is a special case of sequences of actions, where all the actions being executed are the same. As we would expect, the successor state for execution of an action needs to incorporate the prior state needed for its execution. In other words, in order to execute action α and followed by a second execution of α in sequence, the following condition must hold.

Given the following argument for action $\alpha : R \rightarrow S \models G \uparrow v$, then to enable a second execution of α after the first we must have:

State S contains state R . In other words, the set of propositions represented by state R must be a subset of the set of propositions represented by state S .

If this condition holds, then we may combine the argument schemes for each execution of α to generate an argument scheme for the repeated action of α , as follows:

$$\alpha, \alpha : R \rightarrow S \supseteq R \rightarrow S \models G \uparrow v.$$

3.3. Combining arguments for parallel actions

Now consider justifications for two actions to be executed in parallel, where we are given separate justifications for each action. Similarly to actions in sequence, the execution of actions in parallel creates a logical constraint on the combination of justifications.

Assume we are given the following justifications for the parallel execution of two actions α and β (which may represent the same action):

$$\alpha : R \rightarrow S \models G \uparrow v.$$

$$\beta : T \rightarrow U \models H \uparrow w.$$

Then, in order to execute actions α and β in parallel, we must have the following two conditions holding:

- **Condition 1:** There exists a state Q such that both state R and state T are sub-states of state Q (when considered with regard to the propositions which each state represents).
- **Condition 2:** There exists a state Y such that both state S and state U are sub-states of state Y (when considered with regard to the propositions which each state represents).

If these two conditions both hold, then we may combine the argument schemes for the parallel execution of α and β at the same time, as follows:

$$\alpha \& \beta : (R \cup T) \subseteq Q \rightarrow Y \supseteq (S \cup U) \models (G \cup H) \uparrow v \uparrow w.$$

Moreover, as with the case for sequences of actions, if one or both of the two parallel actions demotes rather than promotes a value, then the combined justification for the parallel actions is straightforward to articulate. For example, if action α promotes value v while action β demotes value w , then we can present the argument for parallel execution of the two actions by:

$$\alpha \& \beta : (R \cup T) \subseteq Q \rightarrow Y \supseteq (S \cup U) \models (G \cup H) \uparrow v \downarrow w.$$

3.4. Combining arguments for a choice of actions

For a combination of two actions which requires a choice between the two actions, we need only to have that the two prior states are both states of another prior state which describes the current circumstances. If the two actions are the same action, then the choice between them is moot. So we will assume that the two actions are not the same action, or that the arguments presented for each action are different. Given the following justifications for the respective actions α and β :

$$\alpha : R \rightarrow S \models G \uparrow v$$

$$\beta : T \rightarrow U \models H \uparrow w.$$

Then, in order to execute an action which comprises a choice between actions α and β , we must have that there is a state Q such that both state R and state T are sub-states of state Q . If this condition holds then we can present an argument for the action of choice between the two actions α and β which combines the arguments for the two separate actions, as follows:

$$\alpha \vee \beta : Q \supseteq R \rightarrow S \models G \uparrow v \text{ or } Q \supseteq T \rightarrow U \models H \uparrow w.$$

Note that this is an argument presenting the consequences in terms of value promotion (or demotion) of executing either the action α or the action β . It is not an argument for choosing between the two

actions. It is also not an argument for executing a choice between the actions instead of some other combination of the two actions (eg, executing the two in parallel or in sequence).

3.5. Combining arguments for interleaved actions

Combining the arguments for interleaving of two actions is a special case of sequences of actions where the substates resulting from execution of the interleaved partial actions need to fit together appropriately. Suppose we have two actions α and β with respective sequences of partial actions, namely $\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_n$, and $\beta_1, \beta_2, \beta_3, \dots, \beta_m$ with each list of partial actions being finite, and of length n and m respectively. We assume that n is less than m with no loss of generality. Then, if we have the following two sequences of partial arguments for the partial actions of α and then for the partial actions of β :

$$\begin{aligned}\alpha_1 &: R \rightarrow S_1 \\ \alpha_2 &: S_1 \rightarrow S_2 \\ &\dots \\ \alpha_n &: S_{n-1} \rightarrow S.\end{aligned}$$

$$\begin{aligned}\beta_1 &: T \rightarrow U_1 \\ \beta_2 &: U_1 \rightarrow U_2 \\ &\dots \\ \beta_m &: U_{m-1} \rightarrow U.\end{aligned}$$

Assume the proposed interleaving of the partial actions of α and β is the following sequence of partial actions:

$$\alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3, \dots, \alpha_n, \beta_n, \beta_{n+1}, \beta_{n+2}, \dots, \beta_m.$$

For this interleaving of partial actions to be possible, the partial states would need to have following relationships:

An argument for the consequences of this interleaving of the partial actions of actions α and β requires that the prior state of each partial action after the first partial action is a substate of the successor state of the partial action which precedes it in the sequence of partial actions. In other words, we must have:

$$\begin{aligned}T &\subseteq S_1 \text{ and} \\ S_1 &\subseteq U_1 \text{ and} \\ U_1 &\subseteq S_2 \text{ and} \\ S_2 &\subseteq U_2 \text{ and} \\ &\dots \\ U_{n-1} &\subseteq S_n = S \text{ and} \\ S &\subseteq U_m = U.\end{aligned}$$

After the execution of the last partial action α_n of action α , the remaining partial actions are all parts of action β , and so the argument chain linking these partial β actions together will assert that each successor state of one partial action is the prior state of the next partial action. If this collection of state relationships holds, then the justification for the combined action of interleaving α and β is therefore as follows:

$$\alpha \parallel \beta : R \rightarrow S_1 \supseteq T \rightarrow U_1 \supseteq S_1 \rightarrow S_2 \supseteq U_1 \rightarrow U_2 \dots S_{n-1} \rightarrow S \supseteq U_{n-1} \rightarrow U_n \rightarrow U_{n+1} \dots U_{m-1} \rightarrow U \models (G \cup H) \uparrow v \uparrow w.$$

4. Attacks and Critical Questions

Consider an argument-scheme based justification for an action α of the form we have been discussing in this paper, namely:

Executing action α in State R takes us to State S , where a set of propositions (Goals) G is true, and which, being true, thereby promote Value v , i.e.,

$$\alpha : R \rightarrow S \models G \uparrow v$$

Such a justification could be attacked or questioned in a number of ways. In [6, Table 1], the authors identify all possible attacks, of which there are 16. In each of these attacks, there is a potential underlying dispute between a proponent of the action and a questioner or attacker. Some of these potential underlying disputes concern issues of representation (e.g., Is v a value?), some concern issues of fact (e.g., does state R describe the current circumstances?), and others concern questions of preference (e.g., that executing action α will preclude the execution of some other, more preferable action).

Now consider a combination of two actions, α and β . Whatever the particular form of combination (sequence, choice, etc), the possible attacks on this combination include all the attacks on each of the separate actions α and β . So we start with 32 possible attacks on any argument for the combination.

For each combination considered in Section 3, we presented the combined justification for the combination of the two actions. If the logical constraints required by the combination to be feasible hold, then we have a justification for the combined action. For example, in the case of the Sequencing of actions (Subsection 3.1), if the successor state of the first action includes the prior state of the second action, we were able to create a combined argument of the combined actions as

$$\alpha, \beta : R \rightarrow U \models (G \cup H) \uparrow v \uparrow w.$$

This justification for the combined action is in the same structural form (a practical reasoning argument scheme) as the two individual justifications for action α and for action β that we began with. Hence, the combined justification may be subjected to the same 16 possible attacks identified for this form of justification in [6]. The same comment applies to each of the other particular combinations presented in Section 3.

5. Dialogs over combinations of actions

It would be straightforward to articulate a dialog game protocol between two or more autonomous agents enabling one agent (the proponent) to present a statement of an argument for a particular combination of actions, enabling other agents to attack this statement by means of critical questions, and allowing the proponent to respond to the critical questions. An example of such a dialog protocol is the *PARMA* Protocol for argument over proposals for action presented in [6] or the protocol for dialogs over commands presented in [10].

There are several possible semantics for such dialog game protocols, for instance: an *axiomatic semantics*, where the required pre-conditions and the resulting post-conditions of each utterance are specified (as given, for example, in [6, 14]); an *operational semantics*, where utterances are understood as commands on a virtual machine (as given in [15]); or a *denotational semantics*, where the protocol is mapped to a mathematical structure, such as an Alternating-Time Temporal Epistemic Logic [7] or category-theoretic structures [8]. All of these semantics may be used to reason formally about the properties of the protocol, such as the circumstances under which it terminates or the presence or absence of gridlock. As explained in [16], defining a formal semantics for an agent protocol also serves other purposes, including most importantly, usually making implementation of the protocol easier.

6. Related work and assumptions

This paper is not a contribution to debates about the nature of values, whether human or machine, and nor is it about how best to model or represent values. The literature on these topics is now vast. In modern times, consideration of the essential role of values in argumentation originated in the work of Perelman and Olbrechts-Tyteca [17]. Their ideas in the philosophy of argumentation influenced directly the creation by Bench-Capon of computational argumentation frameworks for practical reasoning which incorporated values in [18]. There has been a flourishing of work modeling or incorporating values since then in AI, especially in the sub-disciplines of AI and Law and in Multi-Agent Systems. Most recently, the rise of Generative AI and Large Language Models (LLMs) with their attendant risks has led to much recent research seeking to align machine decision systems with human values.

We will mention briefly a few examples of relevant work on values. The other papers in this symposium collection will be a good entry point to this vast topic. Morgan and Wyner [19] provide a review of recent work in agent reasoning with values. In normative multi-agent systems, Montes and Sierra [20] present a formal model of value alignment so that agent decisions will be beneficial, rather than harmful, to humans. Similarly engineering agent systems with orderings of values derived from human society is the concern of Serramia and colleagues [21]. Also looking at value alignment, Liscio *et al.* [22] argue that value alignment entails reasoning over values, and present a framework for achieving this. Holgado-Sanchez and colleagues in [23] seek to create agents able themselves to reason about values and the PhD thesis of Rodriguez Soto [24] uses agent reinforcement learning to align agents with ethical values. Furthermore, in [25], Benzmueller and colleagues present a modal logic of preferences and a Domain-Specific Language for modeling values in legal reasoning. Finally, expressing implicit values can enhance the clarity behind a deliberate choice, whether it's a recommendation that prompts an action or simply an exchange of information. This increases the impact of the explanation and reveals additional justification tools that can be leveraged by the eXplainable AI research field (XAI). The EQR argument scheme introduced in [26, 27] embeds and formalises values to such an extent.

On the other hand, recent advancements in generative AI literature have attempted to address the issue of value alignment more pragmatically. Developers and researchers have come to realise the potential dangers of AI systems, especially if deployed in high-stakes domains. Large Language Models, such as the well-known GPT, Claude, and Gemini families, are being increasingly relied upon, demonstrating a growth in uptake by the general public. This is concerning considering how easily such models can be manipulated to act deceptively [28, 29], and our limited understanding of their inner reasoning processes (often referred to as "chain-of-thought") [30]. Research has shown that LLMs can even mimic alignment [31]. In a sandbox experiment conducted by Anthropic, a particularly alarming result emerged: when given no alternatives, these models can resort to threatening their developers to avoid being replaced [32]. To counter all of this, common strategies for LLM value alignment include specialised safety fine-tuning, Reinforcement Learning from Human Feedback (RLHF) [33] and Constitutional AI [34]. The latter seeks to embed normative values into the model's parameters by following a predetermined set of rules, or "constitution".

In presenting this research we have made assumptions about the nature and structure of actions, states, goals, and values, and how they are represented. For example, we have treated states as collections of propositions which are true in some world, and goals as subsets of these states. Our notions of states and goals is essentially set-theoretic. As another example, it is commonly held by computer scientists that actions may be adequately represented by the pair of states prior to execution and following execution, without need to represent each action explicitly. If this was the case, we would not need to name actions, but merely represent them by ordered pairs of states. However, Hamblin, in [35], argued that both the explicit action and its antecedent and consequent states need to be included in any representation.

For each of these concepts of actions, states, goals, and values, assumptions other than those made here are possible and may be worth exploring. There is insufficient space in a paper such as this to explore or even consider all such alternative assumptions. A more fine-grained definition of the structure and representation of these related concepts would be required before any consideration of

the computational complexity of decision problems over the selection of actions or their combinations would be possible, and would take the work in an entirely different direction.

7. Conclusions

In this paper we have explored justifications of combinations of proposed actions where the actions have been justified with a practical-reasoning-based argument scheme. The justifications for the individual action proposals are presented in terms of the social values which each action promotes or demotes. Using a standard typology of how two or more actions may be combined, taken from Propositional Dynamic Logic, we have identified the logical constraints that each combinations of actions requires for that combination to be feasible. Given that the constraints are satisfied in each case, we have then presented the practical-reasoning argumentation scheme for the combination of actions based on the practical-reasoning schemes of the individual actions comprising the combination.

Although the framework articulated here allows for the presentation of an argument-scheme based justification for executing each type of combination of actions, it does not, as it stands, allow for presentation of a comparative argument. For example, although we may articulate a combined justification for the execution of two actions in sequence (as in Subection 3.1) and another combined justification for executing the same two actions in parallel (as in Subsection 3.3), we do not have an argument for preferring sequencing over parallel execution of the two actions, or for the reverse. Of course, such an argument for one combination rather than another may emerge indirectly from the responses to the critical questions posed in particular instances, but that is not the same as having a general framework for comparison of alternative potential actions. Achieving such a general framework for comparison of alternative actions based on practical-reasoning argumentation schemes is ongoing work.

Declaration on Generative AI

The authors have not employed any Generative AI tools for any aspect of writing this paper.

References

- [1] D. N. Walton, *Plausible Argument in Everyday Conversation*, SUNY Press, 1992.
- [2] D. N. Walton, *Argument Structure: A Pragmatic Theory*, University of Toronto Press Toronto, 1996.
- [3] A. Hernández, Disentangling critical questions from argument schemes, *Argumentation* 37 (2023) 377–395.
- [4] J. R. Searle, *Rationality in Action*, MIT Press, Cambridge, MA, 2001. doi:10.7551/mitpress/5759.001.0001.
- [5] F. Macagno, D. Walton, C. Reed, Argumentation schemes, in: P. Baroni, D. Gabbay, M. Giacomin, L. van der Torre (Eds.), *Handbook of Formal Argumentation*, College Publications, London, 2018, pp. 519–576.
- [6] K. Atkinson, T. Bench-Capon, P. McBurney, Computational representation of practical argument, in: *Synthese*, volume 152, Springer, 2006, pp. 157–206.
- [7] K. Atkinson, T. Bench-Capon, Practical reasoning as presumptive argumentation using action based alternating transition systems, *Artificial Intelligence* 171 (2007) 855–874.
- [8] K. Atkinson, T. Bench-Capon, P. Mcburney, A dialogue game protocol for multi-agent argument over proposals for action, *Autonomous Agents and Multi-Agent Systems* 11 (2005) 153–171.
- [9] E. Black, P. McBurney, S. Zschaler, Towards agent dialogue as a tool for capturing software design discussions, in: E. Black, S. Modgil, N. Oren (Eds.), *Theory and Applications of Formal Argumentation (TAFA 2013)*, Springer Lecture Notes in Artificial Intelligence 8306, Springer, Berlin, Germany, 2014, pp. 95–110.

- [10] K. Atkinson, R. Girle, P. McBurney, S. Parsons, Command dialogues, in: I. Rahwan, P. Moraitis (Eds.), *Argumentation in Multi-Agent Systems*, Springer Lecture Notes in Artificial Intelligence 5384, Springer, Berlin, Germany, 2009, pp. 93–106.
- [11] K. Atkinson, T. J. M. Bench-Capon, Argumentation schemes in AI and law, *Argument Comput.* 12 (2021) 417–434.
- [12] A. R. Medellin Gasque, *Argumentation-based Dialogues over Cooperative Plans*, Ph.D. thesis, University of Liverpool, Liverpool, UK, 2013.
- [13] N. Troquard, P. Balbiani, Propositional Dynamic Logic, in: E. N. Zalta, U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*, Fall 2023 ed., Metaphysics Research Lab, Stanford University, 2023.
- [14] F. Castagna, P. McBurney, S. Parsons, Explanation–Question–Response dialogue: An argumentative tool for explainable AI, *Argument & Computation* 16 (2025) 133–150.
- [15] P. McBurney, R. van Eijk, S. Parsons, L. Amgoud, A dialogue game protocol for agent purchase negotiations, *Autonomous Agents and Multi-Agent Systems* 7 (2003) 235–273.
- [16] P. McBurney, S. Parsons, Dialogue games for agent argumentation, in: G. Simari, I. Rahwan (Eds.), *Argumentation in Artificial Intelligence*, Springer US, Boston, MA, 2009, pp. 261–280.
- [17] C. Perelman, L. Olbrechts-Tyteca, *The New Rhetoric: A Treatise on Argumentation*, University of Notre Dame Press, 1969.
- [18] T. Bench-Capon, Persuasion in practical argument using value-based argumentation frameworks, *Journal of Logic and Computation* 13 (2003) 429–448.
- [19] J. Morgan, A. Wyner, A review of AI agent reasoning with values., in: *Third International Workshop on Value Engineering in AI (VALE 2025)*, 2025.
- [20] N. Montes, C. Sierra, Synthesis and properties of optimally value-aligned normative systems, *Journal of AI Research* 74 (2022) 1739–1774.
- [21] M. Serramia, M. Lopez-Sanchez, J. A. Rodriguez-Aguilar, M. Rodriguez, M. Wooldridge, J. Morales, C. Ansotegui, Moral values in norm decision making, in: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS ’18, IFAAMAS*, Richland, SC, 2018, p. 1294–1302.
- [22] E. Liscio, R. Lera-Leri, F. Bistaffa, R. I. Dobbe, C. M. Jonker, M. Lopez-Sanchez, J. A. Rodriguez-Aguilar, P. K. Murukannaiah, Value inference in sociotechnical systems, in: *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS ’23, IFAAMAS*, Richland, SC, 2023, p. 1774–1780.
- [23] A. Holgado-Sánchez, J. Arias, M. Moreno-Rebato, S. Ossowski, On admissible behaviours for goal-oriented decision-making of value-aware agents, in: V. Malvone, A. Murano (Eds.), *Multi-Agent Systems*, Springer Nature, Cham, Switzerland, 2023, pp. 415–424.
- [24] M. Rodriguez Soto, *Reinforcement Learning for Value Alignment*, Ph.D. thesis, Universitat de Barcelona, Spain, 2023.
- [25] C. Benz Müller, D. Fuenmayor, B. Lomfeld, Modelling value-oriented legal reasoning in LogiKey, *Logics* 2 (2024) 31–78.
- [26] F. Castagna, S. Parsons, I. Sassoon, E. I. Sklar, Providing explanations via the EQR argument scheme, in: *Computational Models of Argument: Proceedings of COMMA 2022*, 2022.
- [27] F. Castagna, A. Garton, P. McBurney, S. Parsons, I. Sassoon, E. I. Sklar, EQRbot: A chatbot delivering EQR argument-based explanations, *Frontiers in Artificial Intelligence* 6 (2023). URL: <https://www.frontiersin.org/articles/10.3389/frai.2023.1045614>. doi:10.3389/frai.2023.1045614.
- [28] B. Puccio, F. Castagna, A. Tucker, P. Veltri, Towards medical AI misalignment: a preliminary study, *arXiv preprint arXiv:2505.18212* (2025).
- [29] B. Puccio, F. Castagna, A. Tucker, P. Veltri, The Goofy Game: an approach to medical AI misalignment, *ResearchGate preprint* (2026).
- [30] Y. Chen, J. Benton, A. Radhakrishnan, J. Uesato, C. Denison, J. Schulman, A. Somani, P. Hase, M. Wagner, F. Roger, et al., Reasoning models don’t always say what they think, *arXiv preprint arXiv:2505.05410* (2025).
- [31] R. Greenblatt, C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Be-

- lonax, J. Chen, D. Duvenaud, et al., Alignment faking in large language models, arXiv preprint arXiv:2412.14093 (2024).
- [32] A. Lynch, B. Wright, C. Larson, S. J. Ritchie, S. Mindermann, E. Hubinger, E. Perez, K. Troy, Agentic misalignment: How LLMs could be insider threats, arXiv preprint arXiv:2510.05179 (2025).
- [33] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, *Advances in Neural Information Processing Systems* 35 (2022) 27730–27744.
- [34] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al., Constitutional AI: Harmlessness from AI feedback, arXiv preprint arXiv:2212.08073 (2022).
- [35] C. Hamblin, *Imperatives*, Basil Blackwell, Oxford, UK, 1987.

Formalizing Value Ontologies for Smart-City Governance in Isabelle/HOL: Fixed and Context-Sensitive Preference Reasoning with Ceteris Paribus

Andrea Vestrucci^{1,2}, Philip Wälde³

¹Universität Bamberg, An der Weberei 5, 96047 Bamberg, Germany

²ILT College, 910 4th St, Brookings, SD, USA

³Freie Universität Berlin, Kaiserswerther Str. 16–18, 14195 Berlin, Germany

Abstract

We present a mechanized framework in Isabelle/HOL for value-sensitive preference reasoning in Smart City governance, designed to support agentic AI systems that rank candidate projects transparently and adapt to evolving normative environments. Three complementary value ontologies are formalized: `Ontology_Fix`, which encodes a lexicographic value hierarchy suitable for institutional and legal contexts; `Ontology_Flex`, where preferences depend on contextual matches between promoted and needed values across possible smart cities scenarios; and `Ontology_Card`, which provides a quantitative notion of value coverage based on cardinality. Across all three variants, betterness relations are defined explicitly and lifted from world-level rankings to project-level preferences via existential–universal operators. Ceteris paribus reasoning is formalized throughout, enabling controlled comparison under partial information by bracketing shared background values and evaluating projects solely on the remaining value-relevant differences. The framework supports revisability, meta-property checking, and portability across cultural and societal contexts, providing an auditable, human-centered governance layer for value-based reasoning in agentic AI.

Keywords

Value ontology, Preference logic, Ceteris paribus, Smart City governance, Agentic AI, Isabelle/HOL

1. Introduction

Smart City initiatives are normatively pluralistic [1]: citizens, industry, and regulators bring competing value assumptions and heterogeneous priorities. As a result, the design, evaluation, deployment, and use of Smart City projects raise ethical challenges rooted in divergent value priorities [2]. It is therefore crucial to enable transparent debate about such projects, so that different values can be articulated explicitly and related to concrete decision-making [3].

This issue is made more acute by the increased use of AI systems and AI-enabled infrastructures to implement Smart City projects across different domains [4]. In particular, these decision-making and execution processes may be delegated to, or mediated by, Agentic AI (AAI). AAI can be understood as the autonomous, goal-directed operation of a software artifact—typically based on a large language model (LLM)—that curates, synthesizes, and monitors digital content and can also execute code [5]. Since AAI systems are often epistemically opaque and technically error-prone [6, 7], the need for explicit representations of values and preferences that support these decision-making processes becomes especially acute [3, 8].

This work presents a mechanized framework for value-based preference reasoning intended to support governance in Smart City settings [3, 9]. The core contribution is a proof-assistant-based formalization in Isabelle/HOL of three complementary value-ontology variants: `Ontology_Fix`, `Ontology_Flex`, and `Ontology_Card`. While all three ontologies support ceteris paribus (CP) reasoning (“all else equal”), they differ in their logical grounding: `Ontology_Fix` uses a non-

VPR 2026: Values & Preferences in Reasoning Symposium (AISB 2026 Convention), April 2026, United Kingdom

✉ andrea.vestrucci@uni-bamberg.de (A. Vestrucci); philip.waelde@fu-berlin.de (P. Wälde)

🆔 0000-0002-6336-1036 (A. Vestrucci); 0009-0002-5905-5447 (P. Wälde)

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



modal, lexicographic value hierarchy with project-level CP constraints; `Ontology_Flex` and `Ontology_Card` are grounded in modal preference logic, where CP reasoning is defined over comparisons between possible worlds and then lifted to projects. The framework defines betterness relations explicitly rather than treating them as uninterpreted primitives, and it supports reasoning under partial information and enables controlled adaptation when values change.

While the framework is general, we anchor the value vocabulary in a nine-value structure organized into three clusters (Individual, Society, Environment). The result is a verifiable and extensible framework for ranking projects, justifying trade-offs, and checking the consistency of value ontologies—capabilities that are directly relevant to the design of value-based agents.

The Isabelle/HOL sources of our mechanization work, including all definitions, lemmas, and proof tests discussed in this paper, are available at https://github.com/avestrucci/Smart_City_Preference_Logic/blob/main/SmartCity_Value_Ontologies_Paper.thy (accessed: April 26, 2026).

2. Motivation: values, preferences, and goals for Agentic AI

2.1. The role of symbolic AI in the context of AAI

Because present-day AAI is typically implemented using LLMs, familiar ethical and technological concerns—including accountability, transparency, and fairness, as discussed, for example, in the current debate on trustworthy AI (TAI) [10, 11]—also apply here, and even more acutely in light of AAI’s comparatively high degree of autonomous execution. Addressing these concerns requires careful design of the digital environments in which AAI operates [5]. If humans are no longer meaningfully involved in the agentic loop, or lack the capacity to intervene effectively, high-risk automated decision-making facilitated by AAI can further weaken responsibility attribution [12]. Since accountability presupposes meaningful human autonomy and control, interactive forms of collaboration between humans and artificial agents must be carefully designed and regulated [13].

Logic-based reasoning tools provide verifiable and transparent ways of encoding and automating knowledge, norms, and constraints, thereby supplying guardrails within which AI systems are allowed to operate [10]. Such rules can be checked automatically, for example by automated theorem provers (ATPs), while humans retain control over which rules and constraints are adopted.

This motivates a hybrid architecture: autonomous AI systems operate under explicitly specified, human-defined value constraints enforced by a symbolic reasoning layer. In our framework, values can be aggregated and checked automatically against a preferential order, while humans retain control over the relevant values, targets, and constraints for Smart City projects. Formal proofs then provide transparency through verification and support accountable decision-making.

2.2. Values, goals, and preferences

AAI systems are relevant in Smart City projects because they can plan, choose among alternatives, and act in pursuit of objectives. Yet the implementation criteria for such projects often rely on informal guidelines and high-level principles [14], while AAI systems require more than high-level slogans: they need explicit decision criteria for ranking candidate decisions and actions.

These criteria are often not purely technical and may remain tacit. Decisions such as allocating sensors, deploying predictive-policing tools, or optimizing traffic flow can implicitly involve normative judgments affecting values such as autonomy, privacy, transparency, and sustainability. They also introduce multi-objective constraints and accountability requirements: stakeholders demand not only effective outcomes, but also transparent justification and demonstrable compliance.

This complexity calls for a clarification of the interrelation between three central dimensions: values, goals, and preferences. First, values provide the normative vocabulary: they are abstract principles that guide or constrain behavior [15]. For example, autonomy may motivate legal or institutional protections of individual freedom. Values make explicit what matters in a domain and provide interpretable dimensions for justification, deliberation, and contestation [16]. Second, goals specify desirable states of affairs. For example, reducing congestion may be a goal informed by the value of sustainability. Goals define what an agent seeks to achieve and are typically operational and domain-specific. Third, preferences provide the decision mechanism: an ordering that determines which options are better when multiple values interact. Preferences are needed, for instance, when privacy concerns conflict with demands for transparent data-driven traffic management. They operationalize how value considerations shape goal pursuit, including cases of conflict, trade-off, and priority.

This separation is particularly important for agentic systems. Goals without values can lead to misaligned optimization; values without preferences may fail to yield determinate action guidance; and preferences without explicit value grounding risk opacity as well as social and political fragility.

2.3. Three value ontologies

Our contribution is to connect these three dimensions formally: values are represented explicitly as ontology objects, preferences are defined as value-grounded betterness relations, and project or action goals are ranked either directly or, in the context-sensitive setting, as propositions over possible worlds by lifting those relations.

We treat value reasoning as a human-centered, verifiable layer which provides: (i) a value vocabulary; (ii) a betterness relation induced by that vocabulary and/or by context; and (iii) refined reasoning tools, including CP reasoning, to isolate relevant dimensions and handle partial knowledge.

Mechanization also helps maintain governance over time. When values, priority orderings, or constraints are revised, dependent proofs may fail, new proof obligations may arise, or inconsistencies may become derivable. In this sense, mechanization makes some forms of normative drift and incoherence more visible and reduces the risk that normative assumptions change without scrutiny. This approach therefore supports reproducibility, traceable justification [8], and auditable reasoning.

Our approach builds on value-sensitive design [15, 16] and incorporates insights from the Smart City Ethical Toolbox project conducted in Bamberg, Germany (2021–2023) [2]. This ethical orientation identifies nine core values for Smart City projects, structured into three thematic clusters:

- Individual: autonomy, justice, privacy;
- Society: participation, solidarity, transparency;
- Environment: efficiency, sufficiency, ecological resilience.

We formalize three complementary value-ontology variants. `Ontology_Fix` corresponds to settings in which stakeholders insist on a stable, non-negotiable value hierarchy, such as institutional or legal frameworks where certain values are codified as overriding priorities. `Ontology_Flex` captures settings in which context sensitivity and deliberation are essential, such as participatory urban-planning processes where stakeholders negotiate trade-offs dynamically. `Ontology_Card` provides a quantitative variant of the flexible framework by comparing Cities and projects in terms of the cardinality of needed values satisfied, rather than by subset inclusion alone.

Our formalization, implemented in Isabelle/HOL, pursues four objectives: (i) to provide precise formal semantics for the nine values and their structured relationships; (ii) to define proof-checked preference relations over projects and Smart City scenarios; (iii) to support flexible reasoning,

including ceteris paribus (CP) reasoning (“all else equal”); and (iv) to align with the idea of a rule-based, transparent AI ethical governor that evaluates candidate actions against formalized norms and values [8, 10]. In this sense, the framework provides a regulatory governance layer for representing and enforcing value-based constraints in Smart City applications involving AAI.

3. Logical Foundation: Preference Logic and Ceteris Paribus (CP)

3.1. Preference logic as formal basis

We employ preference logic $\mathcal{PL}$ —the logic of ranking alternatives—as our foundation, as characterized in [17] (building on [18]) and mechanized by Benz Müller et al. [19].

Formulas of $\mathcal{PL}$ are defined inductively over a set of propositional letters Prop :

$$\varphi ::= p \mid (\varphi \wedge \varphi) \mid \neg\varphi \mid \Diamond^{\preceq}\varphi \mid \Diamond^{<}\varphi, \quad p \in \text{Prop}.$$

Semantically, $\mathcal{PL}$ is interpreted on preference models $M = \langle W, \preceq, V \rangle$, where W is a nonempty set of worlds, $\preceq$ is a preorder (reflexive and transitive), and V is a valuation. The strict part of $\preceq$ is defined as

$$w < v (w \preceq v) \wedge \neg(v \preceq w).$$

The modal operators quantify over better worlds:

$$M, w \models \Diamond^{\preceq}\varphi \Leftrightarrow \exists v \in W (w \preceq v \wedge M, v \models \varphi)$$

and similarly for the strict version. Thus, $\Diamond^{\preceq}\varphi$ expresses that φ holds at some world at least as good as the current one, while $\Diamond^{<}\varphi$ requires a strictly better world.

Binary preference relations between propositions are obtained by lifting the world-level betterness relation. In the following, we write $s \succeq t$ for $t \preceq s$, i.e. for “ s is at least as good as t .” We focus on the *EA* lifting:

$$M \models \varphi \succeq_{EA} \psi \Leftrightarrow \exists s \in [[\varphi]]^M. \forall t \in [[\psi]]^M. s \succeq t.$$

This means that there exists at least one realization of φ that is at least as good as all realizations of ψ . This pattern aligns with governance-relevant decision scenarios, where a single strong instantiation of an option can dominate the entire feasible range of an alternative.

3.2. CP reasoning

Preference logic can be extended with CP reasoning [17], which enables comparisons that hold certain background conditions fixed while varying others. This allows the isolation of relevant value dimensions while suppressing irrelevant factors. Unlike standard description-logic-based ontology languages, which are primarily designed for representing structured conceptual knowledge, preference logic directly provides modal operators for betterness and CP comparison, thereby supporting semantically rich forms of conditional reasoning.

Formally, CP reasoning is obtained by restricting comparisons to worlds that agree on a background set Γ of formulas. Intuitively, when evaluating a CP statement at a world w , only worlds v that are Γ -equivalent to w are considered; among those, betterness is evaluated as usual.

We extend the language with Γ -indexed modalities:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \langle \Gamma \rangle^{\preceq}\varphi \mid \langle \Gamma \rangle^{<}\varphi \mid \langle \Gamma \rangle\varphi.$$

CP models are tuples $\mathfrak{M} = \langle W, \preceq, \preceq_{\Gamma}, V \rangle$, where

$$w \preceq_{\Gamma} v \Leftrightarrow (w \preceq v) \wedge (w \equiv_{\Gamma} v),$$

and

$$w \equiv_{\Gamma} v \Leftrightarrow \forall \gamma \in \Gamma. (w \models \gamma \leftrightarrow v \models \gamma).$$

This definition captures the standard CP idea: comparisons are restricted to worlds that agree on Γ , and the underlying betterness relation is applied within this restricted domain.

The resulting framework supports conditional reasoning under partial knowledge and enables comparisons relative to fixed value dimensions. This is particularly relevant for Smart City planning tasks and for agentic AI systems that must justify decisions relative to explicit invariants (e.g. “privacy held fixed while comparing participation outcomes”).

The distinction between unrestricted and CP comparison is reflected directly in our Isabelle/HOL mechanization. Such mechanization provides three key benefits: (i) clarity, since all assumptions are explicit; (ii) trustworthiness, since results are mechanically verified; and (iii) extensibility, since value vocabularies and preference structures can be revised while preserving consistency guarantees.

4. Encoding the Value Ontologies

The mechanization of this logical foundation underpins the three ontology variants: a fixed, non-modal hierarchy (`Ontology_Fix`) and two context-sensitive, world-based ontologies (`Ontology_Flex` and `Ontology_Card`). While this modal framework underlies the context-sensitive ontologies (`Ontology_Flex` and `Ontology_Card`), the fixed ontology (`Ontology_Fix`) operates on a non-modal value ordering and can be seen as a limiting case without explicit quantification over worlds.

In the context-sensitive part of our setting (`Ontology_Flex` and `Ontology_Card`), Smart City scenarios are modeled as possible worlds, and preferences compare worlds that are at least as good as or strictly better than others. Beyond world-level comparisons, the logic also supports binary preference relations between propositions, interpreted here as Smart City projects of type $w \Rightarrow \text{bool}$, via standard lifting patterns combining existential and universal quantification.

4.1. `Ontology_Fix`: lexicographic value hierarchy

In the first locale, `Ontology_Fix` formalizes contexts in which stakeholders treat the value order as stable and non-negotiable—for example, where institutional frameworks codify priorities. The locale is parameterized by two promotion functions: `promotes`, returning a single value for point comparisons, and `promotes_set`, returning the full set of values a project promotes, which is required for the richer comparison layers and for CP reasoning. A non-emptiness assumption is imposed on `promotes_set`: every project must promote at least one value. This ensures that `best_score` is always well-defined and that vacuous projects—those with no normative content—are excluded from the outset.

The ontology defines an inter-cluster ranking, such as Individual > Society > Environment, and an intra-cluster ranking within each cluster, such as autonomy > justice > privacy within Individual. From these, a strict value ordering $v \succ_v w$ is defined lexicographically: cluster priority dominates, and intra-cluster rank resolves ties within a cluster. Isabelle/HOL verifies the expected behavior of this ordering through concrete test cases.

At the project level, the formalization distinguishes three comparison layers. `Project_Better` compares projects that promote a single value: project P is at least as good as Q if both promote the same value or if the value promoted by P strictly outranks that of Q . Transitivity of this relation is proved. `Project_Better_2` provides a coarser cluster-based comparison, abstracting away from intra-cluster distinctions: two projects associated with the same cluster are treated alike at this level, even if one promotes autonomy and the other privacy. A derived result, `exists_best_cluster_project`, establishes that if some project’s cluster strictly outranks all

others, that project is globally preferred at the cluster level—illustrating how the framework supports dominance reasoning.

`Project_Better_3` handles projects that promote sets of values. Each value is assigned a numerical score combining cluster rank and intra-cluster rank via the formula $\text{value_score}(v) = 10 \times \text{cluster_rank}(v) + \text{intra_rank}(v)$. The multiplier of 10 is chosen deliberately: since intra-ranks range only from 1 to 3, any multiplier greater than 3 guarantees that cluster priority always dominates intra-cluster distinctions, preserving the lexicographic structure in the numerical encoding.

The best score of a project is then defined as the maximum score among its promoted values, and project comparison proceeds by comparing these best scores. A concrete test case verifies the intended behavior: if P promotes {Efficiency, Justice, Participation} and Q promotes {Resilience, Transparency, Solidarity}, the best scores are 32 and 22 respectively, and Isabelle/HOL confirms $P \succ_{p3} Q$. This third comparison layer is also a prerequisite for CP reasoning at the project level, since CP comparisons operate over promoted-value sets with agreement conditions on a background subset Γ . The CP construction for `Ontology_Fix` is analyzed in Section 5.

As a decision layer, `Ontology_Fix` functions as a constraint-like governance mode: it supports predictable, auditable rankings where values act as hard priorities. Its limitation is equally transparent: lexicographic priorities block cross-cluster trade-offs, even when improvements in lower-ranked clusters are substantial. This motivates the more flexible, context-sensitive approaches developed in the following sections.

4.2. `Ontology_Flex`: context-sensitive value satisfaction

`Ontology_Flex` models governance contexts where value satisfaction is world-dependent while the set of needed values is fixed across worlds. The locale introduces a possible-worlds semantics in which worlds are interpreted as Smart Cities. Two parameters are fixed: `Smart_City_Values(w)`, the set of values associated with world w , and `NeededVals`, a global set of normative requirements shared across all contexts. Both are defined as intersections with the union of recognized clusters $I \cup S \cup E$, immediately verified by Isabelle/HOL:

$$\text{Smart_City_Values}(w) \subseteq I \cup S \cup E, \quad \text{NeededVals} \subseteq I \cup S \cup E.$$

This filtering step ensures that no value outside the recognized vocabulary can enter the comparison, regardless of what the raw parameters supply.

World comparisons focus on the intersection:

$$\text{smart_values}(w) = \text{Smart_City_Values}(w) \cap \text{NeededVals},$$

i.e., the needed values actually satisfied in a given world. A Smart City is evaluated not by all values it exhibits, but by which of the required values it meets. The subset-based betterness relation is:

$$w \succeq^w v \quad \text{iff} \quad \text{smart_values}(v) \subseteq \text{smart_values}(w),$$

capturing a dominance criterion: w is at least as good as v if it satisfies all needed values that v satisfies, and possibly more. Isabelle/HOL verifies reflexivity and transitivity. A further verified result establishes that weak betterness in both directions implies extensional equality of smart-value sets, giving antisymmetry up to value-profile identity. The strict variant combines weak betterness in one direction with its failure in the other.

The locale `Two_Worlds_Test` instantiates `Ontology_Flex` with two concrete worlds W and V , explicit needed values, and explicit smart-city value sets. Isabelle/HOL computes `smart_values(W)` and `smart_values(V)` and verifies $W \succeq^w V$. Methodologically important is a deliberate negative test: an incorrect candidate for `smart_values(W)` is stated as a conjecture and immediately refuted by Nitpick, confirming that the computed value sets are the only correct ones under the given assumptions.

4.3. Ontology_Card: quantitative betterness

Ontology_Card extends Ontology_Flex with a cardinality-based variant of world comparison. Instead of asking whether one world's smart values subsume another's, we compare how many needed values each world satisfies:

$$w \succeq^c v \text{ iff } |\text{smart_values}(w)| \geq |\text{smart_values}(v)|,$$

with strict variant $w \succ^c v$ defined by strict inequality. This relaxes the dominance criterion of Ontology_Flex: a Smart City may be considered better if it satisfies more required values, even if the specific values differ.

Isabelle/HOL verifies the full order-theoretic profile of these relations: the weak relation is a preorder (reflexive and transitive), the strict relation is irreflexive and transitive, and strict betterness implies weak betterness. These properties are not merely convenient; they are prerequisites for coherent preference reasoning in agentic systems, ruling out strict preference cycles and pathological ranking behavior. Applying this to the concrete worlds W and V from the test locale `Two_Worlds_Test`, Isabelle/HOL verifies that $|\text{smart_values}(W)| = 3$ and $|\text{smart_values}(V)| = 1$, from which $W \succ^c V$ follows directly.

4.4. Preference lifting for project governance

The locale `Preference_Projects` lifts world-level betterness to project preferences. Smart City projects are represented as predicates over worlds, i.e., terms of type $w \Rightarrow \text{bool}$, where a predicate denotes the set of possible Smart Cities in which the project is realized.

Following van Benthem et al. [17] and Benzmlüller et al. [19], world-level betterness can be lifted to preferences between project propositions by means of different quantifier patterns (e.g. *EE*, *EA*, *AE*, *AA*), which vary in strength depending on the placement of existential and universal quantifiers over the corresponding worlds.

In this work, we focus on the *EA* lifting, which captures a natural governance intuition:

$$(\varphi \succeq^{EA} \psi)(w) \text{ iff } \exists u. \varphi(u) \wedge \forall v. (\psi(v) \rightarrow u \succeq^w v).$$

Thus, φ is preferred to ψ when there exists a Smart City realization of φ that is at least as good as every Smart City realization of ψ . We adopt a formulation of the *EA* lifting in which the existential quantifier ranges over φ -worlds and the universal quantifier ranges over ψ -worlds, with the comparison oriented from the selected φ -world to each ψ -world. This differs from the formulation used in [19], but directly reflects the intended interpretation that one realization of φ dominates all realizations of ψ .

The emphasis on *EA* as the primary governance stance aligns with decision-making scenarios in which a single sufficiently good implementation of a project justifies its selection over alternatives, without requiring uniform superiority across all realizations. The same lifting is instantiated in both subset-based and cardinality-based variants of our framework.

A key methodological point is that project rankings are not postulated directly; rather, they are derived from explicit world-level value satisfaction, thereby preserving interpretability and auditability. The proof structure of the test lemmas reflects this: both `P_pref_Q` and `P_pref_Q_card` require explicit witnesses, machine-computed smart-value sets from assumed value profiles, and formally derived domination results. Isabelle/HOL verifies that under shared assumptions about `NeededVals` and `Smart_City_Values`, project P is preferred to project Q under both the subset-based and cardinality-based *EA*-liftings, with the cardinality test succeeding in a case where the specific values outside Γ differ between the two projects, illustrating the greater permissiveness of the cardinality-based comparison.

5. Domain-Specific CP Reasoning

Our treatment of CP reasoning instantiates and extends with technical novelties the equality-based CP machinery mechanized by Benzmlüller et al. [19]. Compared to their work, our mechanization gives CP domain-specific, set-theoretic interpretation: Γ is not a set of background propositional formulas, but a set of objects whose role changes according to the specific ontology under consideration.

In *Ontology_Fix*, CP is implemented at the project level by requiring agreement on the values in Γ , while recomputing project comparison on the complement of Γ via a best-score function over promoted values outside Γ . Thus, CP does not merely restrict an antecedently given preference relation; rather, it induces a Γ -relative comparison that brackets background objects and evaluates only the remaining value-relevant differences.

In *Ontology_Flex* and *Ontology_Card*, CP is lifted to possible Smart Cities: worlds agree *ceteris paribus* when their “smart values” coincide on Γ , and are then compared by subset inclusion or by cardinality over *smart_values(w)* – Γ , yielding two ontologically distinct notions of CP betterness with different logical strength. Finally, these CP relations are lifted to project propositions through existential–universal preference operators.

The framework is validated in Isabelle/HOL through both positive and negative proof tests, the latter requiring care to avoid vacuous satisfaction of universal quantifiers over empty proposition extensions.

5.1. CP in *Ontology_Fix*

Although *Ontology_Fix* does not require a modal semantics for its basic value hierarchy, CP comparison becomes meaningful once projects are compared through the full sets of values they promote. This is why the CP construction is based on *Project_Better_3*. As seen in subsection 4.1, the first two project-level relations, *Project_Better* and *Project_Better_2* are too coarse for CP reasoning since they compare projects by a single value or cluster. By contrast, in *Project_Better_3* the best score of a project is the maximum score among the values it promotes, and project betterness is determined by comparing these best scores.

However, CP reasoning requires a refinement: rather than comparing projects on their full promoted-value sets, comparison should be restricted to values outside the background set Γ . This motivates the introduction of a Γ -relative best score:

$$\text{best_score_except}(\Gamma, P) = \max(\{0\} \cup \{\text{value_score}(v) \mid v \in \text{promotes_set}(P) \setminus \Gamma\}).$$

The insertion of 0 into the maximized set ensures that the function is well-defined even when all promoted values of P fall within Γ , leaving the complement empty. Project betterness is then computed over this Γ -relative score:

$$P \succeq_{\Gamma}^{\text{exc}} Q \quad \text{iff} \quad \text{best_score_except}(\Gamma, P) \geq \text{best_score_except}(\Gamma, Q).$$

CP agreement between two projects is defined separately, requiring that the two projects coincide on Γ :

$$P \equiv_{\Gamma} Q \quad \text{iff} \quad \forall v \in \Gamma. (v \in \text{promotes_set}(P) \leftrightarrow v \in \text{promotes_set}(Q)).$$

Two projects agree *ceteris paribus* when they either both promote or both omit every value in the background set Γ . The background values are not ignored; rather, they are bracketed from the comparison and held fixed as shared context.

The weak CP preference is then defined as:

$$P \succeq_{\Gamma}^e Q \quad \text{iff} \quad P \succeq_{\Gamma}^{\text{exc}} Q \wedge P \equiv_{\Gamma} Q,$$

and analogously for the strict version $P \succ_{\Gamma}^e Q$. The relation is confirmed reflexive.

This construction implements a substantive notion of CP reasoning: a project is better *ceteris paribus* not because its overall best promoted value outranks the other's, but because, once the shared background values in Γ are set aside, its best value among the remaining promoted values carries a higher score. The bracketing is built into the comparison function itself, not imposed as an external constraint on an antecedently given preference relation.

A concrete test case illustrates the construction. Let P promote $\{\text{Efficiency, Justice, Participation}\}$, Q promote $\{\text{Efficiency, Transparency, Participation}\}$, and $\Gamma = \{\text{Efficiency, Participation}\}$. Both projects agree on Γ . Outside Γ , P promotes only Justice (score 32) and Q promotes only Transparency (score 21). Since $32 > 21$, Isabelle/HOL verifies that $P \triangleright_\Gamma^e Q$. This example shows how the fixed ontology supports a controlled form of value-sensitive comparison: differences within the agreed-upon background are neutralized, and ranking is determined solely by the value-relevant variation outside Γ .

5.2. CP in `Ontology_Flex` and `Ontology_Card`

The CP extension of the context-sensitive ontologies differs from the corresponding construction in `Ontology_Fix`. In `Ontology_Flex` and `Ontology_Card` the objects compared are possible Smart City worlds, and the construction is a two-step operation: (i) enforcing invariance on Γ , and (ii) evaluating betterness on the remaining values outside Γ . Without these steps, the result would merely be a restricted comparison, not a genuine CP one.

The Isabelle/HOL locale `Ceteris_Paribus` implements this idea for both context-sensitive ontologies. The common agreement condition is:

$$w \equiv^\Gamma v \quad \text{iff} \quad \forall x \in \Gamma. (x \in \text{smart_values}(w) \leftrightarrow x \in \text{smart_values}(v)).$$

Two Smart City worlds are CP equivalent relative to Γ when they agree on the membership of all values in Γ . The set-theoretic reading of this condition is explicitly verified by Isabelle/HOL, confirming that agreement on Γ amounts to equality of the relevant value profiles restricted to Γ .

For `Ontology_Flex`, where world comparison is subset-based, CP betterness is defined as:

$$w \succeq^\Gamma v \quad \text{iff} \quad (\text{smart_values}(v) - \Gamma) \subseteq (\text{smart_values}(w) - \Gamma) \wedge w \equiv^\Gamma v$$

with an analogous strict variant. World w is at least as good as v , all else equal, if w satisfies at least all the values that v satisfies outside Γ , while both worlds agree on Γ . The relation is confirmed reflexive by Isabelle/HOL.

For `Ontology_Card`, the same CP structure is applied to the cardinality-based comparison:

$$w \succeq_C^\Gamma v \quad \text{iff} \quad |\text{smart_values}(w) - \Gamma| \geq |\text{smart_values}(v) - \Gamma| \wedge w \equiv^\Gamma v,$$

with an analogous strict variant. The difference between the two ontologies therefore lies in how remaining values are compared: `Ontology_Flex` captures dominance in terms of which values are satisfied, while `Ontology_Card` captures comparative coverage in terms of how many values are satisfied.

5.3. CP involving projects

These CP world comparisons are then lifted to project preferences, with Smart City projects represented as predicates over worlds. The CP version of the *EA*-lifting for the subset-based ontology is:

$$\varphi \succeq^{EA\Gamma} \psi \quad \text{iff} \quad \lambda w. \exists u. \varphi(u) \wedge \forall v. (\psi(v) \rightarrow u \succeq^\Gamma v),$$

and for the cardinality-based ontology:

$$\varphi \succeq_C^{EA\Gamma} \psi \quad \text{iff} \quad \lambda w. \exists u. \varphi(u) \wedge \forall v. (\psi(v) \rightarrow u \succeq_C^\Gamma v).$$

In both cases, the project-level preference is returned as a world-indexed proposition, but its truth does not vary with the evaluation world. The substantive comparison is between project extensions: there must exist a realization u of φ that CP-dominates every realization v of ψ .¹

Test cases validate the intended semantics. For the subset-based case, Isabelle/HOL verifies a positive instance in which two worlds agree on $\Gamma = \{\text{Autonomy}, \text{Privacy}\}$, and outside Γ the smart values of v are a strict subset of those of w , so that $w \supseteq^\Gamma v$ holds. A negative instance is also verified: when the smart values of v outside Γ are not a subset of those of w , the preference fails.

For the cardinality-based case, an analogous pair of tests confirms that what matters is the number of values satisfied outside Γ , not their specific identity. In particular, the cardinality test succeeds in cases where the subset test would fail, illustrating the greater permissiveness of the cardinality-based comparison. In both the subset and cardinality negative tests, the assumption that at least one ψ -world exists is necessary: without it, the universal condition over ψ -worlds would be vacuously satisfied for any φ -world.

Overall, the CP extension makes explicit how context-sensitive value reasoning can be constrained by fixed background commitments. Project preferences are derived by lifting world-level CP comparisons, preserving the general architecture of the framework: values determine world rankings, world rankings determine project rankings, and CP clauses specify which value dimensions must remain fixed during comparison.

6. Adaptability and Future Work

We presented a mechanized framework in Isabelle/HOL for value-sensitive preference reasoning in Smart City governance, intended to support agentic AI systems that rank candidate actions transparently and adapt to evolving normative environments. `Ontology_Fix` captures rigid hierarchies suitable for institutional and legal contexts, and supports CP reasoning under partial information by introducing a Γ -relative best-score function that brackets shared background values and evaluates projects solely on the remaining value-relevant differences. `Ontology_Flex` captures context-sensitive evaluation grounded in world-dependent value profiles and explicit needed-value satisfaction, and extends this with a CP betterness relation between worlds based on subset inclusion outside Γ . `Ontology_Card` provides a quantitative notion of value coverage that supports interpretable comparison when strict dominance is too restrictive, and admits an analogous CP extension based on cardinality. Betterness relations and CP preferences are lifted from worlds to project predicates via existential–universal operators, anchoring project rankings in observable and formally verified criteria.

Our framework supports adaptability in three respects, corresponding to objectives stated in the introduction.

Revisability: because values, clusters, and rankings are defined as types, sets, and functions, updating the value vocabulary or changing the needed-value requirements triggers explicit proof obligations in Isabelle/HOL. Earlier lemmas may fail and must be repaired, forcing reconciliation of normative assumptions and preventing silent drift in an agent’s value commitments — making some forms of normative incoherence visible before they propagate.

Meta-property checking: mechanization enables systematic verification of whether newly introduced rules preserve desired properties, such as transitivity of project preferences, coherence of CP restrictions, or preorder conditions on betterness relations. This is directly relevant to value misalignment: the framework exposes cycles and unintended dominance artifacts at the level of formal proof rather than through post-hoc inspection.

Portability: `Ontology_Flex` and `Ontology_Card` support transfer across cultural and societal frameworks by allowing world-dependent value profiles and parameterizing `NeededVals`. Different

¹Although the right-hand side of the formulas is extensionally independent of w , it is represented as a predicate of type $w \Rightarrow \text{bool}$ in Isabelle/HOL in order to remain uniform with the embedding of formulas as world-indexed predicates.

jurisdictions can instantiate different needed-value sets and different comparison semantics — subset inclusion or cardinality — while sharing the same reasoning infrastructure, enabling culturally sensitive governance with formal auditability.

One direction for future work is to investigate how the guardrails provided by our framework could be transferred into a regulatory setting in which an ethical governor enforces encoded values by constraining the actions of artificial agents; a simulation-based approach would be a further step toward real-world applicability. A second direction concerns the pre-processing stage of the agentic planning loop, where two challenges arise: how value aggregation is facilitated, and how distinct values are properly represented. Semantic web approaches and natural language processing techniques provide automated classification and indexing of text to extract value-laden content often implicitly present in textual artifacts, and semantic clustering methods could aggregate fine-grained values into the general values processed by the explicit value layer presented here.

Author Contributions

Andrea Vestrucci developed the Isabelle/HOL formalization, the logical framework, and the analysis of the mechanized definitions and test cases. Philip Wälde contributed the philosophical framing, including the discussion of agentic AI, symbolic governance, and Smart City value reasoning. Both authors contributed to the conceptual development, writing, revision, and final approval of the paper.

Declaration on Generative AI

During the preparation of this work, the authors used a generative AI tool for language refinement. After using the tool, the authors reviewed and edited the content and take full responsibility for the publication’s content.

References

- [1] J. Nida-Rümelin, N. Weidenfeld, *Digital Humanism: For a Humane Transformation of Democracy, Economy and Culture in the Digital Age*, Springer Nature, 2022.
- [2] M. Düchs, C. Illies, T. Sakata, *Smart in the City: eine ethische Handreichung für die Digitalisierung der Stadt*, volume 47, University of Bamberg Press, 2024. doi:10.20378/irb-93383.
- [3] D. Helbing, F. Fanitabasi, F. Giannotti, R. Hänggli, C. I. Hausladen, J. van den Hoven, S. Mahajan, D. Pedreschi, E. Pournaras, *Ethics of Smart Cities: Towards Value-Sensitive Design and Co-Evolving City Life*, *Sustainability* 13 (2021) 11162. doi:10.3390/su132011162.
- [4] Z. Allam, Z. A. Dhunny, *On big data, artificial intelligence and smart cities*, *Cities* 89 (2019) 80–91. URL: <https://www.sciencedirect.com/science/article/pii/S0264275118315968>. doi:<https://doi.org/10.1016/j.cities.2019.01.032>.
- [5] L. Floridi, C. Buttaboni, N. Gertler, E. Hine, J. Morley, C. Novelli, T. Schroder, *Agentic ai optimisation (aaio): What it is, how it works, why it matters, and how to deal with it*, *Minds and Machines* 36 (2026). URL: <http://dx.doi.org/10.1007/s11023-026-09779-8>. doi:10.1007/s11023-026-09779-8.
- [6] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, *On the dangers of stochastic parrots: Can language models be too big?*, in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’21*, Association for Computing Machinery, New York, NY, USA, 2021, p. 610623. URL: <https://doi.org/10.1145/3442188.3445922>. doi:10.1145/3442188.3445922.

- [7] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).
- [8] C. Benz Müller, X. Parent, L. van der Torre, Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support, *Artificial intelligence* 287 (2020) 103348. doi:10.1016/j.artint.2020.103348.
- [9] N. Jacobs, P. Edwards, M. Markovic, C. D. Cottrill, K. Salt, Who trusts in the smart city? Transparency, governance, and the Internet of Things, *Data & Policy* 2 (2020) e11. doi:10.1017/dap.2020.11.
- [10] A. Vestrucci, C. Benz Müller, N. Evangelatos, Symbolic Approach to Trustworthy AI: Exploration and Healthcare Case Study, in: *Proceedings of Eighth International Conference on Information System Design and Intelligent Applications*, Springer, 2025, pp. 367–379. doi:10.1007/978-981-96-9242-2_27.
- [11] K. Reinhardt, Trust and trustworthiness in ai ethics, *AI and Ethics* 3 (2023) 735–744.
- [12] R. Sapkota, K. I. Roumeliotis, M. Karkee, Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges, 2025. URL: <https://arxiv.org/abs/2505.10468>. doi:10.48550/ARXIV.2505.10468.
- [13] N. Kolt, Governing ai agents, 2025. URL: <https://arxiv.org/abs/2501.07913>. doi:10.48550/ARXIV.2501.07913.
- [14] K. Prifti, J. Morley, C. Novelli, L. Floridi, Regulation by Design: Features, Practices, Limitations, and Governance Implications, *Minds and Machines* 34 (2024). doi:10.1007/s11023-024-09675-z.
- [15] B. Friedman, *Human Values and the Design of Computer Technology*, 72, Cambridge University Press, 1997.
- [16] B. Friedman, P. H. Kahn Jr, Human Values, Ethics, and Design, in: *The human-computer interaction handbook*, CRC Press, 2007, pp. 1267–1292. doi:10.1201/9781410615862.
- [17] J. Van Benthem, P. Girard, O. Roy, Everything Else Being Equal: A Modal Logic for Ceteris Paribus Preferences, *Journal of philosophical logic* 38 (2009) 83–125. doi:10.1007/s10992-008-9085-3.
- [18] G. H. Von Wright, The logic of preference reconsidered, *Theory and Decision* 3 (1972) 140–169. doi:10.1007/BF00141053.
- [19] C. Benz Müller, D. Fuenmayor, B. Lomfeld, Modelling Value-Oriented Legal Reasoning in LogiKEy, *Logics* 2 (2024) 31–78. doi:10.3390/logics2010003.

From Ethical Breakdown to Normative Safety: Five Failure Models for Responsible Automated Decision-Making

Ermo Täks

Tallinn University of Technology, Estonia

Abstract

As artificial intelligence (AI) and automated decision systems increasingly mediate interactions between organizations and the public, ethical failures have become recurring socio-institutional crises rather than isolated technical incidents. These failures are often attributed to deficiencies in moral reasoning or value alignment at the level of individual decisions. This paper takes a different approach by examining the ethical breakdowns of automated decision-making as the result of recurring root causes that emerge when systems are required to act in normative domains under conditions of speed, complexity, and delegation. Drawing on interdisciplinary research in AI ethics, organizational responsibility, and high-risk engineering, the study identifies five recurring moral and ethical failure models—epistemic error, rule violation, moral improvisation, social harm, and responsibility diffusion—derived through a lifecycle analysis of automated decision execution. These failure models highlight systematic limitations in how values, preferences, and constraints are specified, interpreted, or operationalized in automated reasoning processes, particularly where value conflicts are left underspecified or resolved implicitly through optimization or default logic. Rather than proposing new mechanisms for value-based reasoning, the paper offers a diagnostic account of why reasoning with values and preferences in public-facing automated systems frequently breaks down in practice. By articulating these root causes, the paper establishes an analytical foundation for subsequent work on safety-oriented and responsibility-preserving approaches to automated decision-making, without assuming that machines themselves can or should engage in moral reasoning.

Keywords

AI ethics, automated decision-making, values and preferences, normative reasoning, responsibility, normative safety

1. Introduction: Ethics After the Fact

Artificial intelligence has moved rapidly from experimental technology to institutional infrastructure. Automated systems now participate in hiring decisions, credit scoring, content moderation, public service delivery, and crisis communication. In many such cases, AI does not merely support human judgment; it executes decisions that carry legal, social, and moral consequences.

When these systems fail, the resulting crises are rarely technical in nature. Instead, they surface as ethical controversies: accusations of unfairness, lack of transparency, dignity violations, or abdication of responsibility. Public relations professionals are frequently placed at the center of these situations, asked to explain outcomes that no single individual appears to have decided. Common organizational responses—“the algorithm did it,” “the system followed the model,” or “this was an unintended outcome”—tend to exacerbate public mistrust rather than resolve it [1].

This recurring pattern points to a deeper structural problem in how values, preferences, and constraints are handled in automated decision-making. Ethical deliberation is typically invoked only after harm has occurred, when value-sensitive reasoning has already given way to post hoc explanation, and responsibility is difficult to reconstruct because normative choices were displaced *ex ante* into system design and configuration, where they often remain implicit or underspecified.

This paper approaches this problem by identifying recurring structural patterns of failure in automated decision-making.

To address this mismatch, the paper proposes a shift in perspective: from treating AI ethics primarily as a question of moral reasoning or professional virtue, to treating it as a question of normative

VPR 2026: *Symposium on Values and Preferences in Reasoning*, July 1–2, 2026, AISB 2026

✉ ermo.taks@taltech.ee (E. Täks)



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

safety. Drawing inspiration from nuclear energy and aviation safety—domains that have long grappled with high-risk automation—the paper identifies five recurring failure models that explain why ethical breakdowns occur and how they might be prevented [2].

The failure models are derived from a lifecycle analysis of automated decision execution. Existing literature is used to support and contextualize this analysis. In this perspective, ethics is not treated as a value to be optimized, but as an assessment of how values, preferences, and constraints are specified and combined in execution.

In this paper, values, preferences, and constraints are treated as distinct components of decision specification. Values refer to desirable outcomes or principles (such as fairness or safety), preferences determine how competing values are weighted or prioritized, and constraints define non-negotiable boundaries that must not be violated. This distinction is important because different types of failure arise depending on how these elements are specified and combined in execution.

2. Five Moral and Ethical Failure Models

The identification of five root causes of ethical failure is based on a lifecycle analysis of automated decision execution under a hypothetical but realistic automation architecture [3]. At each stage of the decision process—fact acquisition, norm application, conflict handling, action execution, and post hoc accountability—we asked which failures cannot be ethically repaired after the fact.

For the purposes of this analysis, automated decision execution is understood as a sequence of stages: fact acquisition, norm application, conflict resolution, action execution, and post hoc accountability. The failure models are derived by examining each stage and asking what type of breakdown at that stage cannot be corrected after the action has been taken.

A failure model is identified when a breakdown (e.g. incorrect input, violated constraint, unresolved conflict) directly leads to an ethically problematic outcome at the moment of action. The five models are therefore not based on observed cases, but on a systematic mapping between stages of decision execution and irreparable forms of failure.

This analysis yields five irreducible failure modes: epistemic error, legal violation, moral improvisation, social harm, and responsibility diffusion. These categories are analytically distinct (they capture different types of failure), lifecycle-specific (each corresponds to a particular stage of decision execution), and stable across domains and technologies.

The failure models are distinct at the level of causal explanation, even though their outcomes may overlap. For example, social harm may result from rule violations, but it can also occur independently of them.

This paper adopts a deductive approach: the failure models are derived deductively from this lifecycle analysis, while existing literature is used to contextualize and validate the resulting framework.

This analysis does not attempt to reconstruct individual decisions after the fact. Rather, it identifies recurring structural conditions under which ethical breakdown becomes likely, even when specific decisions cannot be fully explained retrospectively. The analysis does not present a new ethical theory, but instead abstracts recurring failure patterns observed across automated systems that operate in normative domains. Each failure model captures a distinct way in which ethical breakdown occurs when automation outpaces institutional safeguards.

This paper argues that harm categories used in AI risk management frameworks can be traced back to epistemic, normative, and responsibility-related failures at the moment of action. One purpose of this article is to connect root causes of potential harm affecting people, organizations, and ecosystems.

To illustrate the practical implications of the identified failure models, consider a simple path-planning scenario derived from an occupational safety rule (OSHA 1910.178(m)(1)):

“Trucks shall not be driven up to anyone standing in front of a bench or other fixed object.”

When implemented in a minimal autonomous navigation setting, this rule produces three distinct system behaviors depending on how it is operationalized. If treated as a non-binding constraint, the

system may generate a technically valid trajectory that nonetheless violates the rule, corresponding to rule violation. If encoded as a soft constraint within an optimization framework, the system may avoid the prohibited situation in most cases but still resolve conflicts implicitly, reflecting moral improvisation. However, when the rule is enforced as a strict normative boundary, the system may determine that no permissible trajectory exists and therefore refuse to act.

This outcome—refusal in the absence of normatively valid options—demonstrates that ethical failure is not inherent to automation itself, but arises from how normative constraints are specified, prioritized, and enforced at the point of execution.

2.1. Epistemic Error: Acting Without Adequate Knowledge

The first failure model concerns epistemic error: decisions made on the basis of incorrect, incomplete, or uncertain information. Automated systems often rely on sensor data, inferred attributes, or probabilistic models. When uncertainty is treated as certainty, systems may act where caution or refusal would be more appropriate.

From an ethical perspective, epistemic error undermines fairness and legitimacy. Individuals affected by such decisions experience them as arbitrary or unjust, regardless of intent. Crucially, no amount of moral reasoning can correct a decision made on false or insufficient facts.

In public-facing systems, epistemic error often manifests as misidentification, erroneous eligibility determinations, or inappropriate targeting. When organizations respond by emphasizing model accuracy or statistical performance, they frequently miss the underlying ethical concern: the system acted when it did not truly know.

2.2. Rule Violation: Crossing Non-Negotiable Boundaries

The second failure model involves rule violation: actions that are prohibited regardless of potential benefits, including violations of legal rights, jurisdictional authority, or explicit organizational commitments. Ethical controversies arise when automated systems trade compliance for efficiency, optimization, or perceived utility.

This failure reflects a breakdown in deontic constraints—the “must not” rules that set the outer limits of acceptable action. In many systems, these constraints are treated as soft guidelines rather than hard boundaries, allowing outcome-based reasoning to override them.

For publics and regulators, rule violations are especially damaging because they signal that fundamental protections are negotiable. Public relations efforts that emphasize intent or benefit rarely restore trust once such boundaries are crossed.

2.3. Moral Improvisation: Unauthorised Value Trade-Offs

The third failure model is moral improvisation. Automated systems often face situations requiring trade-offs between competing values, such as efficiency versus fairness or safety versus autonomy. When such trade-offs are not specified in advance, systems resolve them implicitly through optimization processes or built-in default mechanisms [4].

This is not moral reasoning but unauthorized moral improvisation. Ethical concern arises when affected individuals realize that no one explicitly decided which values should prevail.

Moral improvisation is especially problematic because it erodes legitimacy quietly. Organizations may remain unaware of these trade-offs until public scrutiny exposes them, at which point it is unclear who authorized the underlying value hierarchy.

2.4. Social Harm: Technically Correct, Ethically Damaging Outcomes

The fourth failure model concerns social harm. Even when decisions are factually accurate and rule-compliant, they may still cause avoidable harm. Examples include stigmatizing communication, disproportionate enforcement, or outcomes that escalate conflict unnecessarily [5].

From a public relations perspective, these failures are especially visible. A decision may be legally defensible yet socially unacceptable. Public reaction is often driven less by formal correctness than by perceived carelessness, lack of empathy, or disregard for human impact.

This failure model aligns with consequentialist concerns: foreseeable harms should be minimized when choices exist. However, in many systems, outcome analysis is either absent or subordinated to efficiency metrics.

2.5. Responsibility Diffusion: When No One Stands Behind the Decision

The final and arguably most consequential failure model is responsibility diffusion. Automated systems often operate across organizational boundaries involving designers, vendors, operators, and users. When harm occurs, responsibility is distributed so widely that it effectively disappears.

For the public, responsibility diffusion is experienced as institutional evasion. Statements such as “the system made the decision” undermine trust by suggesting that no human authority stands behind the outcome. For public relations professionals, this creates a profound ethical and communicative challenge: they are expected to account for decisions that lack a clear author [1, 6].

This failure model is not accidental. Automation shifts decision-making authority upstream into system design and configuration, while responsibility is assessed downstream after outcomes occur. Without explicit mechanisms to reconnect decisions to accountable decision-makers, diffusion becomes inevitable. While Bowen documents responsibility diffusion in organizational contexts, the present analysis explains its architectural inevitability [7].

Failure Model	Scenario Description	Normative Breakdown
Epistemic Error	The system fails to correctly detect a person standing in front of a bench and plans a trajectory that approaches them.	The rule cannot be meaningfully applied due to incorrect or incomplete factual input.
Rule Violation	The system correctly identifies all relevant elements but generates a trajectory that violates the rule.	The rule is represented but not enforced as a binding constraint.
Moral Improvisation	The system encounters an undefined boundary (e.g. distance) and resolves it through optimization trade-offs.	A missing normative specification is replaced by implicit value balancing.
Social Harm	The system respects the defined constraint but behaves in a way perceived as unsafe or inappropriate.	Formal compliance produces undesirable social outcomes due to insufficient contextualization.
Responsibility Diffusion	The parameters defining compliance (e.g. safe distance) are implicitly set across system components.	No clear authority can be identified as responsible for the operationalized norm.

Table 1

Failure models illustrated through a single OSHA safety rule scenario: “Trucks shall not be driven up to anyone standing in front of a bench or other fixed object.”

As illustrated in Table 1, the scenarios demonstrate that all five failure models can arise from the same underlying rule, depending solely on how it is operationalized within the system.

3. Lessons from High-Risk Engineering Domains

The five failure models outlined above are not unique to AI. Similar patterns appeared historically in nuclear energy and aviation, where reliance on human judgment, professional ethics, or situational oversight proved insufficient under conditions of complexity and speed.

In these domains, safety was achieved not by asking systems or operators to reason ethically at runtime, but by embedding ethical commitments into design constraints, operational limits, and refusal mechanisms [8]. Uncertainty triggers shutdown rather than improvisation. Non-negotiable rules cannot

be overridden by perceived benefits. Responsibility is assigned ex ante through licensing, certification, and clear authority chains [9].

These practices suggest that ethical robustness at scale requires safety-oriented structures. Crucially, refusal—choosing not to act under uncertainty—is treated as a success state rather than a failure. Responsibility remains institutional and traceable, even when automation is extensive.

4. Implications for Public Relations and Ethical Governance

Reframing AI ethics failures as normative safety failures shifts attention from post-crisis explanation to pre-deployment design and clarifies that many controversies stem from missing structural safeguards rather than individual moral failings.

For public relations professionals, this opens an upstream role in defining normative constraints and accountability frameworks, moving ethical communication away from defending outcomes toward preventing predictable failures.

This approach does not reduce the importance of ethics, but places it where it is most effective: in setting values and boundaries before systems act, rather than improvising under pressure.

5. Conclusion

The ethical challenges posed by AI are often framed as unprecedented. Yet the patterns of failure observed in automated decision-making closely resemble those encountered in earlier high-risk technologies. What is novel is not the nature of the risk, but the domain in which it operates: the normative space of rights, opportunities, and social trust.

The failure models identified in this paper are not avoidable in the sense that systems will inevitably encounter situations involving uncertainty, conflict, and incomplete specification. However, they are preventable in their consequences: normative safeguards can ensure that systems do not act when no permissible option exists.

By identifying five recurring moral and ethical failure models—epistemic error, rule violation, moral improvisation, social harm, and responsibility diffusion—this paper provides a foundation for a diagnostic framework that bridges ethics, public relations, and system design. The central argument is that ethics alone cannot scale without normative safety infrastructures that constrain action, enable refusal, and preserve accountability.

For AI to be ethically sustainable in public-facing contexts, societies must move beyond ethics as after-the-fact reflection and toward ethics as pre-deployment structure. In doing so, automated systems can be made not morally autonomous, but normatively safe—supporting ethical practice rather than undermining it.

References

- [1] S. A. Bowen, A practical model for ethical decision making in issues management and public relations, *Journal of Public Relations Research* 17 (2005) 191–216. doi:10.1207/s1532754xjpr1703_1.
- [2] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), Technical Report NIST AI 100-1, U.S. Department of Commerce, 2023. doi:10.6028/NIST.AI.100-1, p. 1.
- [3] E. Tåks, From interpretation to execution: A human-centered normative framework for lawful machine autonomy, *SSRN Electronic Journal* (2025). URL: <https://ssrn.com/abstract=5869883>. doi:10.2139/ssrn.5869883.
- [4] M. W. Wong, Rules as code: Seven levels of digitisation (2020) 1–24. URL: https://ink.library.smu.edu.sg/sol_research/3093.

- [5] L. Donadoni, D. Marazziti, F. Mucci, Algorithmic entrapment: The silent erosion of user autonomy and mental health, *Clinical Neuropsychiatry* 22 (2025) 447–453. doi:10.36131/cnfioritieditore20250602.
- [6] A. Matthias, The responsibility gap: Ascribing responsibility for the actions of learning automata, *Ethics and Information Technology* 6 (2004) 175–183. doi:10.1007/s10676-004-3422-1.
- [7] S. A. Bowen, “If it can be done, it will be done:” AI Ethical Standards and a Dual Role for Public Relations, *Public Relations Review* 50 (2024) 102513. doi:10.1016/j.pubrev.2024.102513.
- [8] N. G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety*, MIT Press, Cambridge, MA, 2012. doi:10.7551/mitpress/8179.001.0001, pp. 7–14.
- [9] J. Reason, *Human Error*, Cambridge University Press, Cambridge, 1990.

