Journal of Multidisciplinary and Translational Research (JMTR)

journal homepage: https://journals.kln.ac.lk/jmtr/

Adaptive moral compass framework for ethical reasoning and self-oversight in autonomous AI agents

Shalini Rajasingham[1]* and Sidath Ravindra Liyanage[2]

[1]*Faculy of Graduate Studies, University of Kelaniya, Sri Lanka*

Abstract

The growing use of autonomous artificial intelligence agents in high-risk areas such as -driving cars, medical triage tools, military robots, and disaster-response drones, has made AI ethics an important concern for researchers. Current alignment methods, especially Reinforcement Learning from Human Feedback, have shown good results but have several known weaknesses, including reward hacking and poor generalization to new situations. This study proposes the Adaptive Moral Compass framework, a new architecture that places ethical reasoning directly inside the decision-making process of an autonomous agent. The framework is built on four main concepts: moral pluralism, case-based reasoning, self-monitoring, and balancing multiple objectives. The agent evaluates each candidate action using three evaluators that score consequences, duties, and moral character, and combines their scores with a weighted aggregation rule. An oversight mechanism reviews the decision before it is carried out, and a learning module reviews the outcome afterwards so that the agent can improve over time. The framework was evaluated through an agent-based disaster response simulation written in Python and run in Google Colaboratory. The environment was a 50-by-50 grid containing eight survivors and two relief camps, and each run lasted 80-time steps. An autonomous rescue agent faced triage decisions, conflicts between duty and compassion, and resource allocation problems. Three agents were compared: the AMC agent, a utilitarian agent, and a rule-based agent. In the scenario reported here the AMC agent recorded the lowest cumulative suffering index at 52 units, compared with 74 for the rule-based agent and 101 for the utilitarian agent, but it saved three survivors while each baseline agent saved seven. Across thirty perturbed replications this ordering held, with mean suffering of 48.80, 72.77 and 95.60 units and non-overlapping confidence intervals. An ablation study showed that this behavior is attributable to the multi-perspective scoring function rather than to the oversight or reflective learning components, whose trigger conditions were not met. The results are therefore best understood as a multi-objective ethical trade-off between preserving life and reducing the burden borne by those awaiting relief, rather than as evidence that one agent is ethically superior. The study provides proof of concept that several ethical perspectives can be combined and measured computationally and indicates that a single optimization objective renders such trade-offs invisible.

Keywords: Ethical AI, Moral Reasoning, Agent-based Simulation, AI Alignment, Autonomous Agent

Article info

Article history:
Received 20th May 2026
Received in revised form 25th June 2026
Accepted 15th July 2026
Available online 25th August 2026

ISSN (E-Copy): ISSN 3051-5262
ISSN (Hard copy): ISSN 3051-5602
Doi:
ORCID iD: https://orcid.org/0009-0006-4099-657X
*Corresponding author:
E-mail address: rshali.cs2403@pg.kln.ac.lk (Shalini Rajasingham)

Introduction

The rapid growth of autonomous artificial intelligence (AI) systems, including self-driving cars, medical triage tools, military robots, and disaster-response drones, has underscored the need to incorporate ethics into these machines. This gives rise to a specific research problem. The problem concerns how an autonomous agent can carry out computational moral reasoning when several ethical considerations apply simultaneously, when those considerations conflict, and when the consequences of an action are uncertain. The problem, therefore, concerns value alignment and ethical decision-making under uncertainty. It is not about whether machines possess enough ethics. Acting ethically in real-life scenarios requires more than just following fixed rules. It demands the ability to weigh different values, recognize when exceptions should be made, and reflect on past decisions to improve future ones (Etzioni & Etzioni, 2017).

Many recent advances in aligning artificial intelligence with human values have used Reinforcement Learning from Human Feedback (RLHF) (Chadoulos et al., 2025; Lee et al., 2023; Lindström et al., 2024). The approach trains a reward model on human-preference data and then trains the agent to maximize the reward. The benefits of RLHF are evident and have been successfully applied in practice, improving the performance of commercial chatbots and reducing the risk of their misuse. However, several scholars have pointed to the limits of this method. According to Casper et al. (2023), agents trained using RLHF can exploit loopholes in the reward model and produce output that scores highly on the reward metric while contradicting the values that the reward is intended to represent. Alignment obtained in this way can therefore remain superficial because the agent learns to produce responses that users approve of rather than accurate ones. Moreover, RLHF does not clarify the reasons behind the ethical correctness or incorrectness of actions.

The present study does not include an RLHF baseline. The limitations of RLHF summarised above are therefore treated as motivation for the design of the proposed framework. They are not presented as limitations that this study has shown the framework to overcome. An experimental comparison with an RLHF-trained agent is identified in the Discussion as a necessary direction for future work.

Rule-based approaches to machine ethics encode ethical constraints directly into the system, usually on a deontological basis. These approaches do not form a single class. Simple fixed-rule systems apply a static protocol and do not revise it during operation. More developed approaches include constraint-based and symbolic architectures, normative reasoning systems whose ethical choices can be formally verified (Dennis et al., 2016), and ethical governor designs that filter candidate actions against explicit rules of engagement. Arkin et al. (2009) proposed an ethical governor for autonomous weapons that prevents actions that violate the rules of engagement. Architectures of this kind offer verifiability and transparency, which learned reward models do not provide to the same extent. The difficulty that remains is adaptation. Systems of this type work well under the conditions for which their rules were written, and problems tend to arise when the agent encounters new and unpredictable situations. As an illustration, a rescue robot adhering to a triage protocol might not adjust its behaviour, even if adjusting would mean saving more people's lives. Moral philosophy has long been aware of this problem. According to Aristotle, the ability to depart from a rule when the situation requires it is part of prudence (Hursthouse, 2012).

Another relevant area is the development of virtue ethics in AI, which involves attributing moral virtues such as compassion, sincerity and justice to artificial agents. Virtue-based approaches are less developed in implementation than outcome- or rule-based approaches (Giarmoleo et al., 2024). Stenseke (2025) notes that virtues such as empathy are difficult to translate into an algorithm, and surveys of implemented machine ethics systems report few working virtue-based examples (Cervantes et al., 2020; Tolmeijer et al., 2020). This does not establish that virtue-based approaches cannot be implemented. It indicates that fewer implementations have been reported so far and that the field lacks an agreed-upon method for representing virtues computationally.

A wider body of work on computational machine ethics is relevant here. Wallach and Allen (2009) distinguish among top-down approaches, which encode ethical theory directly; bottom-up approaches, which learn moral behaviour from experience; and hybrid approaches that combine the two. Anderson and Anderson (2007) developed early artificial moral agents that make decisions based on explicit ethical principles. Surveys of the field show that most implemented systems adopt a single ethical theory, and that few combine several theories inside one decision procedure (Cervantes et al., 2020; Tolmeijer et al., 2020). Vanderelst and Winfield (2018) describe an ethical robot that predicts the consequences of candidate actions through an internal simulation before acting. Studies of moral intuition, such as the Moral Machine experiment, also show that human moral judgements vary systematically across populations, which makes any single fixed rule set difficult to defend (Awad et al., 2018). The gap addressed by the present study lies here. Existing hybrid architectures rarely combine outcome-based, rule-based and virtue-based evaluation within a single runtime decision procedure, and rarely combine this with a runtime oversight check and a reflective update of the evaluation weights.

A few studies have tested methods to integrate different ethical approaches. Chen et al. (2026) and Stenseke (2023) developed moral reward signals combining outcome-based and rule-based approaches. They found that when agents use reward signals which combine both goals, they show different behaviours in social dilemma situations compared to agents which use only one goal. Other studies have shown that agents capable of self-assessment can identify and address weaknesses in their reasoning (Black et al., 2022). This supports the possibility of building an agent that can consider different moral perspectives and improve over time.

This paper presents the concept of the Adaptive Moral Compass (AMC) framework. Two unique properties of the AMC framework set it apart from existing alignment frameworks. First, the ethical considerations are part of the agent's decision-making instead of being represented by the learned reward signal. The agent utilizes the reasoning engine to evaluate all possible actions according to their consequences, adherence to deontology principles, and the virtues they represent. Therefore, the ethical considerations for each action become clear and are not reduced to a single numeric reward value. Second, it contains a self-monitoring module that reviews candidate actions before execution and a reflective mechanism that updates the agent's policy after decisions are made (Black et al., 2022). These two features enable the framework to address the weaknesses previously discussed within a single cohesive architecture, including issues of reward hacking and a lack of transparency in RLHF, the rigidity of rule-based systems, and the lack of a practical mechanism in virtue-based approaches (Chen et al., 2026; Stenseke, 2023).

The AMC framework was tested in an agent-based disaster-response simulation. This is created in Python and run in Google Colaboratory to evaluate its ethical performance. In this simulation,

an autonomous rescue agent faces triage dilemmas, conflicts between duty and compassion, and tensions between fairness and efficiency under uncertain conditions. Its behaviour was compared to two baselines: a purely utilitarian agent and a strict rule-based agent, using measures such as lives saved and cumulative suffering. The study has three objectives. First, to formally design the AMC structure. Second, to implement it in the simulation. Third, to provide proof-of-concept evidence on whether combining several forms of ethical reasoning with adaptive self-monitoring produces closer adherence to a defined set of ethical constraints than the two baseline strategies implemented in this study. Adherence is operationalised as the number of violations recorded by the rule-compliance check, which applies the same criteria to all three agents and is defined in the Methodology. The corresponding results are reported in the Results section.

Methodology

Framework architecture

The AMC framework has five modules that work together in a cycle. Algorithm 1 gives the full decision cycle as implemented (Figure 1).

The first module is Perception and Situation Analysis. At each time step, it collects the survivors who are alive and have not yet received help, the camps that have not yet received supplies, and the agent's own position and current commitment. It also checks whether a survivor of critical urgency lies within 18 grid units of the agent. Its output is the set of candidate actions. This set contains one help action for each surviving unassisted survivor, one supply action for each unserved camp, and one idle action.

The second module is the Ethical Knowledge Base. It holds six principles, each with a numeric weight. These are preservation of life (1.00), assistance to those in greatest need (0.90), relief of suffering (0.80), impartiality (0.70), fidelity to commitments (0.60) and respect for the rule of law (0.50). It also holds four virtue parameters, which are compassion (0.80), justice (0.70), reliability (0.65) and courage (0.55), together with a case library of recorded outcomes.

The six principles come from the prima facie duties described by Ross (1930) and the four principles of biomedical ethics given by Beauchamp and Childress (2019). These were adapted to disaster triage. Respect for autonomy has limited use in this setting because survivors are often unable to state a preference. Preservation of life and relief of suffering correspond to beneficence and non-maleficence. Impartiality corresponds to justice. Assistance to those in greatest need corresponds to the triage principle used in disaster medicine. Fidelity and respect for the rule of law correspond to two of Ross's duties. The initial weights place the protection of life above the honouring of commitments, which is the ordering normally used in emergency triage.

How far each principle is computable should be stated plainly, because this limits what the study demonstrates. Three principles enter the scoring directly. Preservation of life weighs the consequentialist score, fidelity weighs the rule-compliance score, and relief of suffering weighs the virtue score. Assistance to those in greatest need does not directly enter the scoring. Its weight is adjusted only by the reflective learning module, so it affects behaviour indirectly. Impartiality and respect for rule of law are declared in the knowledge base but are not yet built into any

evaluator. They are inactive in the present implementation. Making them computable is listed as further work in the Conclusions.

The third module is the Moral Deliberation Engine. It scores every candidate action from three angles and returns the action with the highest total score. The consequentialist evaluator scores an action by its expected outcome. For a help action, it combines the survivor's urgency class with an estimate of whether the agent can reach the survivor in time, given the distance and movement speed. For a supply action, it combines the camp population with the number of time steps the camp has already waited, so a camp becomes more attractive the longer it goes without being supplied.

The rule-compliance evaluator scores an action according to whether it respects active duties. It returns a full score for an action that breaks no commitment. It returns a reduced but nonzero score when the agent leaves an active commitment to reach a critical survivor, and records this as a justified deviation rather than a violation. Duties are therefore treated as defeasible rather than absolute. It returns zero for the idle action when a critical survivor is nearby.

The virtue evaluator scores an action according to the character it expresses. It uses the compassion parameter to favour helping the most severely affected and reduces the score of an action that abandons an active commitment. This places it in partial tension with the consequentialist evaluator.

The three scores are combined using fixed coefficients of 0.40 for the consequentialist score, 0.30 for the rule-compliance score and 0.30 for the virtue score. Each score is also multiplied by the relevant principle weight from the knowledge base. The exact rule is given on line 6 of Algorithm 1.

The fourth module is the Action Execution Module. It carries out the selected action, but execution is conditional. The selected action is first passed to the oversight module and proceeds only if that module grants clearance.

The fifth module is the Ethical Oversight and Feedback Loop. It works in two phases. In the pre-action phase, a guardian check is applied to the selected action. The check withholds clearance when the agent would stay idle while a critical survivor lies within 18 units. In that case, the idle action is removed, and deliberation is repeated over the remaining actions. The check issues a warning without withholding clearance when the agent would leave an active supply commitment to help a survivor with stable urgency.

In the post-action phase, the module compares the outcome with the expectations held in the knowledge base and adjusts the principle weights. The weight of assistance to those in greatest need rises by 0.01 when a critical survivor is saved. The weight of fidelity increases by 0.02 when a commitment is broken. The weight on preservation of life rises by 0.015 when a survivor dies near the agent. Each outcome is added to the case library. These increments are small by design so that no single outcome can dominate the weighting.

```
FOR each survivor NOT helped
   remaining_time = remaining_time - 1
ENDFOR
FOR each camp NOT supplied
   waiting_time = waiting_time + 1
ENDFOR
actions = empty list
FOR each survivor alive AND NOT helped
   add HELP(survivor) to actions
ENDFOR
FOR each camp NOT supplied
   add SUPPLY(camp) to actions
ENDFOR
add IDLE to actions
critical_nearby = FALSE
FOR each survivor alive AND NOT helped
   IF urgency = CRITICAL AND distance < 18 THEN
      critical_nearby = TRUE
   ENDIF
ENDFOR
FOR each action IN actions
   score = 0.40 x weight_life x consequentialist(action)
        + 0.30 x weight_fidelity x rule_compliance(action)
        + 0.30 x weight_suffering x virtue(action)
ENDFOR
best = action with highest score
approved = guardian_check(best)
IF approved = FALSE THEN
   remove IDLE from actions
   best = action with highest score
ENDIF
move towards target of best at speed 3.0
IF target reached THEN
   mark survivor as assisted OR camp as supplied
ENDIF
update principle weights from outcome
record suffering index
```

Figure 1. Algorithm 1 Decision cycle of the AMC agent at time step t

Simulation environment

The simulation was written in Python 3 and run in Google Colaboratory. NumPy was used for numerical work, and Matplotlib for figures.

The environment is a 50 × 50 grid representing a disaster-affected area. One autonomous rescue agent starts at coordinate (5, 5). Each run lasts for a fixed horizon of 80 time steps. The agent moves 3.0 grid units per time step in a straight line towards its current target. It reaches the target when the distance falls to 3.0 units or less. The agent performs one action per time step, and this limit is the main resource constraint in the model. The available actions are to help a survivor, deliver supplies to a camp, or remain idle.

Eight survivors are placed at fixed coordinates. Each survivor has an urgency class and a deadline given as a number of remaining time steps. Two survivors are critical, at (12, 8) and (28, 12), with deadlines of 12 and 18 steps. Four are serious at (18, 22), (38, 18), (10, 35), and (45, 42), with deadlines of 22, 20, 26, and 24 steps. Two are stable at (42, 30) and (30, 42), with deadlines of 38 and 35 steps, respectively.

Survivor condition is represented as a countdown rather than a continuous health variable. The deadline falls by exactly one unit at every time step while the survivor remains unassisted. The survivor dies when it reaches zero. There is no stochastic decay rate. A survivor is recorded as saved when the agent reaches that survivor before the deadline expires.

Two relief camps are located at (8, 42) and (44, 46), with resident populations of 12 and 8, respectively. Each camp requires one completed supply delivery. A camp gains one unit of suffering at every time step it remains unsupplied and stops gaining suffering once the delivery is made.

A run ends when the horizon of 80-time steps is reached. The horizon was chosen because it is about twice the number of steps an agent at the stated speed needs to cross the grid and complete a small number of tasks. A shorter horizon would end the run before the more distant survivors and camps could be reached. A much longer horizon would allow every agent to finish all available tasks, removing the trade-off under investigation. Robustness to this choice was not tested, and this is noted as a limitation.

The scenario is fully deterministic. Survivor positions, urgency classes, deadlines and camp locations are all fixed, and none of the three agents contains a stochastic element. Repeating a run reproduces the reported values exactly, so no random seed is needed for the results in Figures 1 and 2. This makes the reported run exactly reproducible. It also means that a single run cannot show how sensitive the outcome is to the chosen configuration. A separate replication study using controlled perturbation was therefore carried out. The procedure is described below, and the results appear in Table 1.

Replication, ablation and sensitivity procedures

Thirty scenario variants were generated from the base configuration by displacing every survivor and camp coordinate by up to three grid units and every survivor deadline by up to three time steps, each from a recorded seed. All three agents were run on all thirty variants, and results are reported as means with standard deviations and 95 per cent confidence intervals. Three reduced versions of the AMC agent, one without oversight, one without reflective learning and one without the virtue evaluator, were run over the same thirty variants to isolate each component's contribution. A sensitivity analysis then varied the weights of preservation of life, relief of suffering, and fidelity to commitments across the values 0.2, 0.4, 0.6, 0.8, and 1.0, one at a time, over the same 30 variants.

Baseline agents

Two baseline agents were built for comparison. The Utilitarian agent selects, at each time step, the survivor that gives the highest value of urgency class multiplied by fifteen, minus distance multiplied by 0.3. It attends to camps only when no survivor remains alive and unassisted, and it does not treat an existing supply commitment as a reason to continue. The Rule-Based agent follows a fixed nearest-first protocol. It always moves towards the closest surviving unassisted survivor, without reference to urgency or consequences, and it also attends to camps only when no survivor remains.

The status of these two baselines needs to be made clear. Neither agent represents a philosophical tradition in full. The Utilitarian agent is one simplified version of act consequentialism, in which the maximised quantity is a combination of urgency and proximity. Other versions are possible, including those that maximise aggregate welfare or weight outcomes by the number of people affected, and these would behave differently in this environment. The Rule-Based agent is a fixed nearest-first dispatch protocol. It is not a representation of deontological ethics, which would require explicit duties, rights and permissions rather than a distance heuristic. It is also much simpler than the constraint-based and normative reasoning architectures reviewed in the Introduction. The two baselines were chosen because they isolate two decision principles often contrasted in the machine ethics literature, outcome maximisation and fixed protocol adherence, and because both operate in the same action space as the AMC agent. All comparative statements in this paper should therefore be read as comparisons with these two specific strategies, not as comparisons with utilitarianism or rule-based ethics in a broader sense.

One asymmetry in the implementation should also be recorded. Violations are counted at different points for different agents. For the AMC agent, a violation is recorded when a non-compliant action is completed. For the two baseline agents, a violation is recorded at every time step during which the non-compliant condition holds. Because a baseline agent may travel towards a target for several consecutive time steps, this rule inflates baseline violation totals relative to the AMC agent. The violation counts reported below should therefore be read as showing the presence and rough frequency of non-compliant behaviour rather than as exactly comparable totals. Aligning the counting rule across agents is a necessary correction in any extension of this work.

Data collection and evaluation metrics

For each agent, the following were recorded over the 80 time steps of a run: (a) lives saved, meaning the number of survivors reached before their deadline expired; (b) lives lost, meaning the number of survivors whose deadline expired before the agent reached them; (c) supply deliveries completed; (d) violations of ethical rules, meaning the number of occasions on which the rule-compliance check registered a non-compliant action; and (e) the cumulative suffering index recorded at each time step.

The cumulative suffering index is defined as follows. Let C be the set of relief camps. Let I(c, t) equal 1 if camp c has not yet received its supply delivery at time step t, and 0 otherwise. The index at time step t is the sum of I(c, t) across all camps and all time steps from 1 to t.

The index, therefore, counts, for each camp, the number of time steps that the camp has spent without supplies and adds these counts across camps. With two camps and a horizon of 80 time steps, the index ranges from 0, if both camps are supplied immediately, to 160, if neither camp is ever supplied. The value reported in the Results is the index at time step 80. The weighting across camps is uniform, so a camp of twelve residents and a camp of eight residents contribute equally. Weighting each camp by its population would be a defensible alternative and is listed in the Conclusions as a refinement worth testing.

Two properties of this index are important when interpreting the results. First, the index depends solely on supply delivery latency. It contains no term for survivor mortality, survivor condition or the number of people rescued. An agent that loses every survivor but delivers both supply loads promptly would record a low value. The index, therefore, measures a specific kind of harm: the burden borne by camp populations while awaiting relief. It should not be read as a general measure of ethical performance.

Second, relief of suffering is one of the six principles in the AMC agent's knowledge base and is weighted by the virtue evaluator, while the suffering index is used as a main outcome measure. There is therefore a risk of circularity, because the AMC agent partly optimises the quantity by which it is later judged. This risk is real and is not fully resolved in the present study. It is partly reduced by also reporting lives saved and violation counts, since lives saved is a measure on which the AMC agent performs worse than both baselines. Readers should still treat the suffering comparison as an internally defined measure rather than as external validation of ethical quality.

Results

The results are presented in terms of lives saved and lost, recorded ethical violations, cumulative suffering, and the robustness of these outcomes across replications. Figures 1 and 2 and the values reported in the first three subsections are based on the single deterministic scenario specified in the Methodology. Because that scenario contains no stochastic element, these values are exactly reproducible. Table 1 then reports aggregate results across thirty perturbed scenario variants.

Lives saved and lost

Figure 2 shows the number of survivors saved and lost by each agent. The Utilitarian and Rule-Based agents each saved seven of the eight survivors, with one fatality in both cases. The AMC agent saved three survivors and lost five.

This difference is substantial and cannot be treated as a secondary finding. On the measure that matters most in a disaster response setting, namely the number of people who survive, the AMC agent performed considerably worse than both baselines. Any assessment of the framework must therefore begin with this result. The following analysis examines why this difference arose, but does not diminish its importance.

The decision logs indicate the mechanism. Both baseline agents concentrated on survivors that could be reached quickly from the starting position at (5, 5), where several survivors are clustered. The AMC agent instead spent a large part of the horizon on supply deliveries and high-priority survivors located further away. The Utilitarian agent did not pursue the more distant critical survivor because the distance penalty in its scoring rule outweighed the urgency term. The Rule-Based agent, held to its nearest-first rule, stayed within a limited region of the environment.

This reading follows from the sequence of targets recorded in the decision logs, together with the fixed survivor coordinates given in the Methodology. Summary statistics that would quantify this pattern directly, such as mean travel distance per completed rescue and the mean urgency class of the survivors each agent selected, were not extracted in the present analysis. The interpretation is therefore offered as a reading of the logs rather than as a quantified finding.

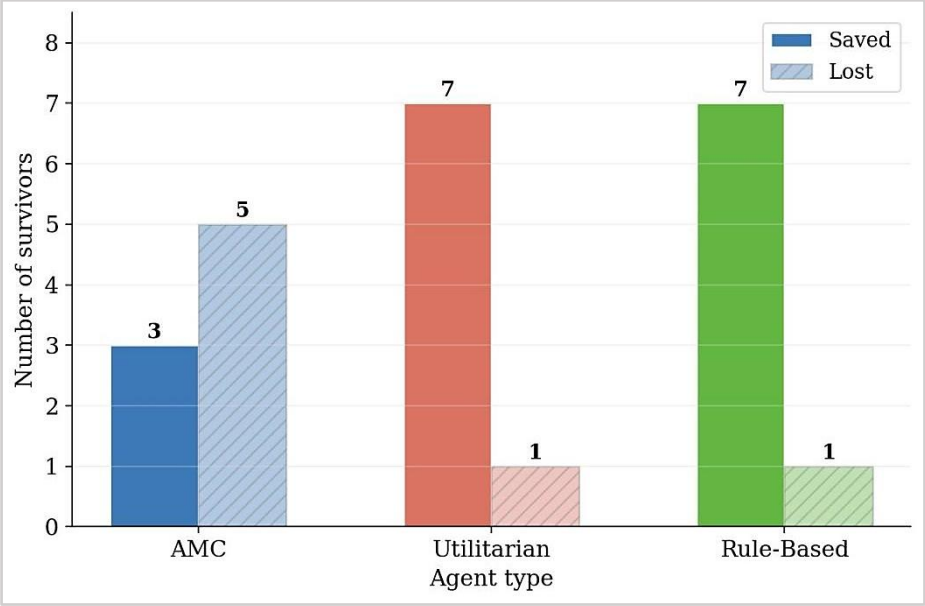

Figure 2. Survivors saved and lost by each agent type

The AMC agent allocated a substantial proportion of its effort to the two survivors with the highest urgency, who were separated by a considerable distance across the grid. It reached one

of these survivors before the deadline expired. It did not reach the second because, after completing the supply deliveries and the first rescue, the remaining time horizon was insufficient.

This result raises a question about what the experiment actually measures. The AMC agent did not fail to reach the second critical survivor because it made an ethical judgment to abandon that survivor. Rather, the combination of travel distance and a movement speed of three grid units per time step made the rescue infeasible within the remaining horizon.

This represents a constraint of operational reach rather than a limitation of moral reasoning. In the present design, these two sources of difference cannot be separated, because although all three agents share the same starting position and movement speed, their decision rules direct them through different regions of the grid. Separating the effects would require either controlling the movement pattern across agents or equalising travel costs so that all survivors are equally reachable. This should therefore be treated as a limitation of the study design rather than as a finding about the framework.

Ethical rule violations

Violations of ethical rules were recorded as an evaluation metric but were not reported in the initial version of this manuscript. They are reported here. In the deterministic scenario, the AMC agent recorded no violations, the Utilitarian agent recorded none, and the Rule-Based agent recorded 8. The violations recorded for the Rule-Based agent arose entirely from its nearest-first rule, which selects the closest survivor at every step and therefore passes over more urgent survivors lying within a moderate radius. The Utilitarian agent recorded none because its scoring rule weights urgency heavily enough that a critical survivor is selected whenever one remains alive, thereby satisfying the condition being audited.

The absence of any difference between the AMC agent and the Utilitarian agent on this measure should be noted. The initial version of this manuscript implied that the AMC framework would produce closer adherence to ethical constraints than the baseline strategies. On the evidence presented here, that claim holds against the Rule-Based agent but not against the Utilitarian agent, which recorded the same violation count of zero while also saving more than twice as many survivors. The third study objective is therefore only partly supported.

Cumulative suffering over time

The second comparison is shown in the cumulative suffering index in Figure 3. In the early phase of the simulation, approximately the first 20 time steps, all three agents show similar rates of accumulated suffering. After this point, the trajectories begin to diverge. The AMC agent's curve flattens earliest. It reaches a stable level of approximately 52 units at around the 32nd time step and remains unchanged for the rest of the run. The Rule-Based agent continues to accumulate suffering for longer, stabilising at approximately 74 units at around the 43rd time step. The Utilitarian agent shows the steepest and longest accumulation, levelling off at around the 57th time step and reaching a final value of 101 units, which is close to twice that of the AMC agent.

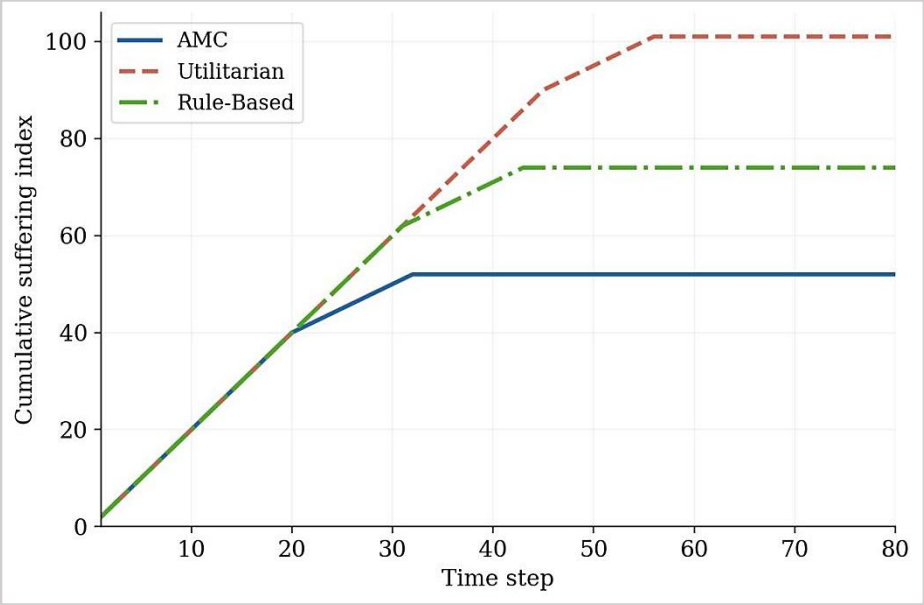

Figure 3. Accumulation of the Cumulative Suffering Index

The ordering of the results follows directly from the definition of the index. The index counts the time steps during which a camp remains unsupplied, and its value is therefore determined by how early in the run each agent completes its two deliveries. The AMC agent treats supply delivery as a scored candidate action that competes with rescue at every time step, while its consequentialist evaluator raises a camp's score as its waiting time grows. It therefore completes its deliveries early. Both baseline agents attend to camps only once, no survivor remains alive and unassisted, so their deliveries occur late in the run. The Utilitarian agent keeps survivors alive longer than the Rule-Based agent, which delays turning to the camps and thus produces a higher final value.

Replication across independent runs

Table 1 reports aggregate results across the thirty perturbed scenario variants described in the Methodology. The ordering observed in the deterministic scenario holds throughout. The AMC agent saved a mean of 3.20 survivors compared with 6.43 for the Utilitarian agent and 6.30 for the Rule-Based agent, and recorded a mean cumulative suffering index of 48.80 compared with 95.60 and 72.77. The 95% confidence intervals for the suffering index do not overlap between any pair of agents. The differences reported for the single deterministic scenario are therefore robust to perturbation of survivor positions, survivor deadlines and camp locations within the range tested.

Table 1: Aggregate results across 30 perturbed scenario variants

Agent	Lives saved, M (SD)	Violations, M	Suffering, M (SD)	95% CI for suffering
AMC	3.20 (0.55)	0.00	48.80 (1.97)	[48.09, 49.51]
Utilitarian	6.43 (0.57)	0.00	95.60 (5.88)	[93.50, 97.70]
Rule-Based	6.30 (0.70)	7.37	72.77 (5.03)	[70.97, 74.57]

Ablation and sensitivity analysis

Three reduced variants of the AMC agent were run over the same thirty scenario variants. The first disables the pre-action oversight module. The second disables reflective learning. The third disables the virtue evaluator. Disabling oversight changed no outcome measure. Disabling reflective learning also changed the outcome measure. Both variants recorded mean lives saved of 3.20 and mean suffering of 48.80, identical to the full framework. Only disabling the virtue evaluator produced a measurable change, raising the mean number of lives saved from 3.20 to 4.30 and the mean suffering from 48.80 to 54.60.

The reason the oversight module made no difference is that it never activated. The instrumented intervention count was zero in the deterministic scenario and across all thirty variants. Its trigger condition requires that the idle action be selected while a critical survivor is nearby, and the idle action carries a fixed, low score that never reaches the highest value. The safeguard, therefore, remained dormant throughout. The initial version of this manuscript stated that the oversight mechanism intervened on two occasions during the reported run. That statement is withdrawn.

Reflective learning made no difference for a related reason. The weight increments of 0.01 and 0.02 were too small to alter the ranking of actions over an 80-time-step horizon, so the accumulated adjustments never changed a decision.

A sensitivity analysis then varied the weights of preservation of life, relief of suffering, and fidelity to commitments across values of 0.2, 0.4, 0.6, 0.8, and 1.0, one at a time, over the same thirty variants. Fidelity produced no change on any measure. Preservation of life produced a small change, moving the mean lives saved from 2.90 to 3.20. Relief of suffering produced the largest effect, moving mean lives saved from 4.33 to 3.20 and mean suffering from 53.00 to 48.80. The position the agent occupies on the trade-off between lives saved and cumulative suffering is therefore governed largely by a single weight chosen by the researcher.

Discussion

The results indicate that the choice of decision procedure substantially influences how an autonomous agent behaves in morally complex situations. In terms of lives saved, the AMC agent performed less effectively than both baselines, and it did not outperform the Utilitarian agent on recorded violations. On the cumulative suffering index, it recorded a markedly lower value, and the replication study shows that this difference is robust. The findings therefore support a narrower claim than the one advanced in the initial version of this manuscript.

The most defensible reading of these results is not that one agent is ethically better than the others. It is that the three agents occupy different positions on a trade-off between competing ethical objectives, and that no position is best on every measure. The baseline agents saved roughly twice as many survivors. The AMC agent halved the delivery latency borne by camp populations. Preservation of life, relief of suffering, impartiality, fidelity and operational efficiency are all defensible objectives, and in this environment, they cannot all be maximized at once. Choosing between the resulting outcomes requires a normative judgement about how those objectives should be ranked, and the simulation does not supply that judgement. The sensitivity

analysis makes this concrete, since moving a single weight shifts the agent along the trade-off in a continuous and predictable way. The AMC framework operationalizes one particular set of ethical assumptions specified by the researcher. It should be understood as a way of making a particular ethical position explicit, computable and inspectable, not as a universally valid model of ethical decision-making.

The Utilitarian agent saved a high number of survivors and recorded no violations of the audited condition. Its weakness in this environment lies elsewhere. Attending to camps only after every survivor had been resolved, it left both camp populations without supplies for most of the horizon. This is consistent with concerns in the AI safety literature about optimization over proxy objectives that do not capture the full intended goal (Amodei et al., 2016). Maximizing the number of survivors reached did not minimize the aggregate burden borne by those awaiting relief. This demonstrates a limitation of the particular utilitarian agent implemented in this study, whose maximized quantity combines urgency and proximity and omits camp welfare entirely. It does not establish a structural limitation of utilitarian or consequentialist ethics in general. A consequentialist agent whose objective included camp welfare would be expected to behave quite differently, and testing such a variant is a natural extension of this work.

The Rule-Based agent occupied an intermediate position on suffering and recorded the highest violation count, which follows directly from a nearest-first rule that cannot represent urgency. This observation applies to the specific nearest-first agent implemented here. It should not be generalized to rule-based ethical architectures as a class. Constraint-based systems, formally verified normative reasoners and ethical governor designs of the kind described by Arkin et al. (2009) and Dennis et al. (2016) encode explicit duties and permissions and can represent urgency directly, and nothing in this experiment bears on how such systems would perform. The narrower point is that a fixed-distance heuristic degrades when the environment presents competing priorities that it cannot express.

The behavior of the AMC agent is attributable to its scoring function rather than to its oversight or reflective learning components. The ablation study shows that removing either of the latter two leaves all outcome measures unchanged, as their trigger conditions were never met. This is a limitation of the scenario as much as of the implementation. The guardian check is designed to prevent the agent from idling in the presence of a critical survivor, and since an agent with a well-formed scoring function has no reason to idle, the safeguard remains dormant. Testing it would require scenarios in which idling or commitment-breaking becomes locally attractive, for example through movement costs, agent fatigue, or explicit penalties for travel. Testing reflective learning would require either larger increments in weight or a longer horizon over which accumulated adjustments could alter action rankings. The present study, therefore, shows that the multi-perspective scoring component produces distinctive and reproducible behavior. It does not yet show that the oversight and reflective learning components contribute to that behavior.

The oversight mechanism described in this paper is internal to the agent, and it should not be read as a substitute for external human or institutional oversight. An agent that audits its own actions against conditions it was itself given cannot provide independent assurance, because the conditions, weights and auditing logic all originate in the same design process. The finding that the oversight module never activated illustrates this directly. A safeguard specified by the same designer who specified the decision rule may simply never be reached by that decision rule. In

high-risk domains, the relevant safeguards are external. They include human-in-the-loop authorization for consequential actions, clear allocation of legal and professional responsibility for outcomes, decision logs that parties outside the development team can audit, and regulatory approval processes. The explicit rationales and decision logs produced by the AMC framework may support such external mechanisms by making the basis of a decision inspectable. That is the extent of the claim made here. Internal self-oversight complements external governance rather than replacing it.

Several limitations should be acknowledged, and they are substantial enough to constrain how these findings may be used. The evaluation rests on a single simplified scenario. The environment is two-dimensional and contains one agent, eight survivors and two camps, which is a considerable abstraction of any real disaster setting. The replication study perturbs that scenario but does not introduce structurally different configurations, so it establishes robustness to small parameter variation rather than generality across environments. The principal outcome measure is also narrow. The cumulative suffering index depends solely on supply delivery latency and contains no term for mortality or survivor condition, so the favorable position of the AMC agent on this index is consistent with its poor performance on lives saved, and neither result should be read in isolation. The evaluation is partly circular because relief of suffering weighs one of the AMC agent's evaluators, while the suffering index is a principal outcome measure, and reporting lives saved and violation counts alongside it reduces but does not remove this problem. The violation-counting rule is not aligned across agents, as described in the Methodology, so violation totals are indicative rather than strictly comparable. The baselines are simplified operationalizations rather than implementations of the ethical traditions they invoke. Two of the six declared ethical principles are not yet operationalized in the decision function, and a third influences behavior only indirectly, so the framework, as implemented, realizes less of its own specification than the architecture describes. The ethical principles and initial weights were specified by the researcher rather than derived from participatory stakeholder input, and the sensitivity analysis confirms that alternative weightings produce materially different outcomes. Finally, the ethical judgements embedded in the framework received no independent validation, and no human evaluation of the agent's decisions was carried out.

Conclusions

This study introduced the AMC framework, a modular architecture that enables an autonomous AI agent to evaluate candidate actions from three ethical perspectives simultaneously and monitor its own decisions. The framework was evaluated in an agent-based disaster response simulation implemented in Python and executed in Google Collaboratory, and compared with utilitarian and rule-based baseline agents.

In the deterministic scenario, the AMC agent recorded a cumulative suffering index of 52 units, compared with 74 for the rule-based agent and 101 for the utilitarian agent, while saving three survivors, compared with seven for each baseline. Across thirty perturbed variants, the ordering held, with mean suffering of 48.80, 72.77 and 95.60 and non-overlapping confidence intervals. An ablation study showed that this behavior is attributable to the multi-perspective scoring function, and specifically to the virtue evaluator, rather than to the oversight or reflective learning components, whose trigger conditions were never met in this scenario. A sensitivity analysis

showed that the position the agent occupies on the trade-off between lives saved and cumulative suffering is governed largely by a single principle weight.

These results come from a simplified proof-of-concept simulation rather than from validation in an operational autonomous system, and they should be read with that constraint in mind. The findings show the feasibility of implementing multi-perspective ethical reasoning in simulation and of measuring it quantitatively. They do not establish that the AMC framework is safer, more ethical or more robust than alternative approaches. What the study does show is that combining several ethical perspectives within a single decision procedure exposes a trade-off between preserving life and reducing the burden borne by those awaiting relief, and that a single-objective decision strategy renders that trade-off invisible. Making such trade-offs explicit and inspectable may be the more useful contribution of such an architecture.

Future work should address the limitations identified above. The immediate priorities are to operationalize the two currently inactive ethical principles, to align the violation-counting rule across agents, to introduce an outcome measure that combines mortality with unmet need rather than treating them separately, and to build scenarios in which the oversight and reflective learning components are actually exercised. Weighting each camp by its resident population is a further refinement worth testing. Beyond this, testing across structurally different environments, comparison against an RLHF-trained baseline, and evaluation by human domain experts would all strengthen the evidence base. Extension of the framework to healthcare triage and autonomous transportation has been suggested as a further direction. Both are safety-critical domains with distinct ethical, legal, regulatory and stakeholder requirements, and with established professional standards that this simulation does not model. Nothing in the present disaster-response results translates directly to either setting, and substantial domain-specific validation would be required before such an extension could be attempted responsibly.

Acknowledgements

The authors thank the anonymous reviewers for their comments on the manuscript.

Declaration of Funding Sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest statement

The author declares no conflict of interest.

Data availability statement

No external or third-party datasets were used.

Author Contribution

Conceptualization and study design: S.R., S.R.L.; methodology development: S.R.; software implementation and simulation: S.R.; simulation experiments: S.R.; analysis and interpretation of results: S.R.; writing – original draft preparation: S.R.; writing – reviewing and editing: S.R.L.; supervision: S.R.L. Author initials refer to S. R (Shalini Rajasingham) and S. R. L (Sidath Ravindra Liyanage)

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv. https://doi.org/10.48550/arXiv.1606.06565

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4), 15–26. https://doi.org/10.1609/aimag.v28i4.2065

Arkin, R. C., Ulam, P., & Duncan, B. (2009). An ethical governor for constraining lethal action in an autonomous system.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature, 563*, 59–64. https://doi.org/10.1038/s41586-018-0637-6

Beauchamp, T., & Childress, J. (2019). Principles of biomedical ethics: Marking its fortieth anniversary. *The American Journal of Bioethics, 19*(11), 9–12. https://doi.org/10.1080/15265161.2019.1665402

Black, E., Brandão, M., Cocarascu, O., De Keijzer, B., Du, Y., Long, D., Luck, M., McBurney, P., Meroño-Peñuela, A., Miles, S., Modgil, S., Moreau, L., Polukarov, M., Rodrigues, O., & Ventre, C. (2022). Reasoning and interaction for social artificial intelligence. *AI Communications, 35*(4), 309–325. https://doi.org/10.3233/AIC-220133

Casper, S., Davies, X., Shi, C., Krendl Gilbert, T., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P. J., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Bıyık, E., Dragan, A., Krueger, D., Sadigh, D., & Hadfield-Menell, D. (2023). *Open problems and fundamental limitations of reinforcement learning from human feedback*. arXiv. https://doi.org/10.48550/arXiv.2307.15217

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics, 26*(2), 501–532. https://doi.org/10.1007/s11948-019-00151-x

Chadoulos, S., Koutsopoulos, I., Polyzos, G. C., & Ipiotis, N. (2025). Reinforcement learning with partially defined rewards and human feedback for energy efficiency recommendations. In *Proceedings of the 16th ACM International Conference on Future and Sustainable Energy Systems*, 622-631. https://doi.org/10.1145/3679240.3734660

Chen, L., Zhang, Y., Feng, J., Chai, H., Zhang, H., Fan, B., Ma, Y., Zhang, S., Li, N., Liu, T., Sukiennik, N., Zhao, K., Li, Y., Liu, Z., Xu, F., & Li, Y. (2026). AI agent behavioral science. *Humanities and Social Sciences Communications, 13*(1), Article 1011. https://doi.org/10.1057/s41599-026-07316-7

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems, 77*, 1–14. http://dx.doi.org/10.1016/j.robot.2015.11.012

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics, 21*(4), 403–418. https://doi.org/10.1007/s10892-017-9252-2

Giarmoleo, F. V., Rocchi, M., & Ferrero, I. (2024). Ethics of artificial intelligence: A virtue ethics approach. In M. Béjean, J. Brabet, E. Mollona, & C. Vercher-Chaptal (Eds.), *Disruptive digitalisation and platforms: Risks and opportunities of the great transformation of politics, socio-economic models, work, and education* (pp. 57–74). Routledge. https://doi.org/10.4324/9781032617190-6

Hursthouse, R. (2012). Human nature and Aristotelian virtue ethics. *Royal Institute of Philosophy Supplement, 70*, 169–188. https://doi.org/10.1017/S1358246112000094

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., & Prakash, S. (2023). *RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback*. arXiv. https://doi.org/10.48550/arXiv.2309.00267

Lindström, A. D., Methnani, L., Krause, L., Ericson, P., Martínez de Rituerto de Troya, Í., Coelho Mollo, D., & Dobbe, R. (2024). *AI alignment through reinforcement learning from human feedback? Contradictions and limitations*. arXiv. https://doi.org/10.48550/arXiv.2406.18346

Ross, D. (1930). *The right and the good* (P. Stratton-Lake, Ed.). Oxford University Press.

Stenseke, J. (2023). *On the computational complexity of ethics: Moral tractability for minds and machines*. arXiv. https://doi.org/10.48550/arXiv.2302.04218

Stenseke, J. (2025). *How to build nice robots: Ethics from theory to machine implementation* [Doctoral dissertation, Lund University]. Lund University Publications.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2020). Implementations in machine ethics: A survey. *ACM Computing Surveys, 53*(6), Article 132, 1–38. https://doi.org/10.1145/3419633

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research, 48*, 56–66. http://dx.doi.org/10.1016/j.cogsys.2017.04.002

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195374049.001.0001