ME: Modelling Ethical Values for Value Alignment

Eryn Rigley[1], **Adriane Chapman** [1], **Christine Evers** [1], **Will Mcneill** [1]

[1]University of Southampton, United Kingdom
e.rigley@soton.ac.uk, Adriane.Chapman@soton.ac.uk, C.Evers@soton.ac.uk, will.mcneill@soton.ac.uk

Abstract

Value alignment, at the intersection of moral philosophy and AI safety, is dedicated to ensuring that artificially intelligent (AI) systems align with a certain set of values. One challenge facing value alignment researchers is accurately translating these values into a machine readable format. In the case of reinforcement learning (RL), a popular method within value alignment, this requires designing a reward function which accurately defines the value of all state-action pairs. It is common for programmers to hand-set and manually tune these values. In this paper, we examine the challenges of hand-programming values into reward functions for value alignment, and propose mathematical models as an alternative grounding for reward function design in ethical scenarios. Experimental results demonstrate that our modelled-ethics approach offers a more consistent alternative and outperforms our hand-programmed reward functions.

Code — https://github.com/erynrigley/ME-Modelling-Ethical-Values-for-Value-Alignment

Introduction

As artificial intelligence (AI) systems are adopted for a variety of high-impact applications with increasing autonomy, they stand to bring about morally concerning and dangerous results. These threats have catalysed research efforts in 'value alignment' which, broadly speaking, refers to the alignment of artificial agents with a value, or set of values (Ji et al. 2024; Gabriel 2020). A major challenge facing value alignment researchers is in accurately translating abstract moral values into a machine readable format (Hibbard 2014; Brundage 2014).

Reinforcement learning (RL), in which the agent learns independently through exploration, has been used to train AI systems to make decisions which align with the moral preferences of experts (e.g., Peschl 2021), programmers (e.g., Abel, MacGlashan, and Littman 2016), and lay humans (e.g., Noothigattu et al. 2018) as well as to balance competing ethical and individual objectives (e.g., Rodriguez-Soto et al. 2022). RL based value alignment requires defining a reward function which promotes ethical behaviours and penalises unethical behaviours. Case-based rewards, wherein the agent receives a manually defined value given some conditional outcome as a results of its actions, are a simple approach (Abouelazm, Michel, and Zoellner 2024). For example, a reward function for a medical delivery unmanned air vehicle (UAV) may punish 1 minute delays by -10. But, why should the punishment be -10 and not -50, and will this punishment work in different environments? Though crude, this example reflects the reality and challenge of programmers manually tuning real number case-based reward functions in line with the size and complexity of the environment (Gupta et al. 2023; Knox et al. 2023; Booth et al. 2023). At the same time, across many domains, humans do not make ethical decisions based solely on their own intuitions, but make use of evidence and advancements in data availability and modelling. For this reason, we turn to domain-specific models as an alternative to hand-set reward functions.

Several novel contributions are made to the field of value alignment in this paper. First, we examine the so far overlooked challenges involved in designing effective reward functions for RL based value alignment. Second, we introduce the use of mathematical models, accepted and used within the domains that AI systems will be deployed into, as an alternative to capturing ethical values for reward functions. We compare the effectiveness of hand-programmed and modelled-ethics approaches to reward function design for aligned RL systems in three related gridworld experiments. Third, we focus on three branches of ethics rarely applied within value alignment research but which are directly applicable to kinds of scenarios in which AI systems will be deployed: disaster ethics, ecocentrism, and healthcare ethics.

Background

Bottom-up approaches to value alignment assume that machines can independently learn how to act ethically if they receive enough correctly labelled input data and computing power, employing techniques such as artificial neural networks and RL (Tolmeijer et al. 2020). RL has been an area of particular interest to value alignment researchers, and is argued to be the most promising, well-suited, approach, in allowing the system freedom and flexibility to discover ethically optimal actions (Kaas 2021; Vishwanath, Dennis, and Slavkovik 2024).

Reinforcement Learning

RL defines a class of algorithms which solve problems modelled as Markov Decision Processes (MDP). MDPs are defined as a tuple, (S, A, P, R, γ), where S is a set of possible states, A is a set of possible actions, P is a probability function which denotes the probability of moving from one state to another given a certain action, R is a reward function specifying a reward as a consequence of being in a state and taking an action, and γ is a discount factor which specifies the present value of future rewards, where a reward received at t time steps in the future is worth only γ^{t-1} times what it would be worth if it were received immediately (Wu and Lin 2017; Ng and Russell 2000; Sutton and Barto 2018). As standard, we define the value of being in state s and taking a action a to end up in state s' as the expected return of starting from s, taking action a, and thereafter following policy π (Sutton and Barto 2018; Metz and Bukov 2023, S.6).

Q-learning was an early breakthrough in reinforcement learning (Watkins and Dayan 1992), and is commonly used as a framework for value alignment (e.g., Rodriguez-Soto et al. 2022; Ecoffet and Joel 2021; Wu and Lin 2017). Here, the learned action-value function, Q, directly approximates the optimal action-value function, $Q*$, so that correct convergence can be achieved if all pairs continue to be updated (Sutton and Barto 2018). In line with related RL approaches to value alignment, we employ a standard Q-learning algorithm, defined as

$$\begin{aligned} Q_{k+1}(s,a) &\leftarrow Q_k(s,a) + \alpha\delta_k, \\ \delta_k &= r(s,a) + \gamma \max_{a'} Q_k(s',a') - Q(s,a) \end{aligned} \tag{1}$$

where k denotes the iteration step of the algorithm, α is the learning rate, and δ_k is the temporal difference error (Watkins and Dayan 1992; Sutton and Barto 2018; Metz and Bukov 2023, S.8). We use this classical approach to RL since our primary focus is not on advancing RL frameworks (such as deep reinforcement learning), but on the technical and moral challenges involved in translating ethical considerations into reward functions for aligned RL systems (Trussell 2018). Our RL framework and learning environment is of comparable complexity to a number of related value alignment works (Vishwanath, Dennis, and Slavkovik 2024).

Hand-Setting Normative Reward Functions

RL approaches to value-alignment require a reward function which effectively promotes ethical behaviours and penalises unethical behaviours. Over 10 years ago, Anderson and Anderson (2011, p. 3) argued, "it is important that ethical issues are not left for programmers to decide - either implicitly or explicitly", but rather, we should turn to ethicists and established ethical theories for guidance. And yet, RL based approached to value alignment continue to involve the programmer hand-setting punishments as real values on a single numeric scale, based on their own intuitions (e.g., Abel, MacGlashan, and Littman 2016; Rodriguez-Soto et al. 2023). Hand-setting numeric values through trial and error is typical of reward function design across RL (Booth et al. 2023; Knox et al. 2023). For example, Knox et al. (2023) note a number of papers wherein conditional reward functions for autonomous vehicles return real number punishments, ranging from -10 to -1000 in case of collision. Such case based functions require defining for all possible distinct morally concerning states, s', some corresponding constant value, c:

$$R^h(s,a,s') = \begin{cases} c_1 & s' = \text{morally concerning state}_1, \\ \dots & \dots, \\ c_n & s' = \text{morally concerning state}_n, \end{cases} \tag{2}$$

This is challenging for several reasons. First, it requires defining all distinct morally concerning states. For example: causing a collision; causing a collision involving a child; ...and so on. Second, it requires defining a set of corresponding values for each of these states. For example: -10 if car causes collision, -50 if car causes collision involving a child. Oftentimes, this is achieved through trial-and-error. In Booth et al. (2023)'s experiments, it took experts an average of 4 attempts to effectively design rewards functions for a simple 4x4 gridworld scenario.

The challenges involved in designing effective reward functions are exacerbated by the range of domain-specific applications of AI systems. RL practitioners, who may not be experts in the field, are challenged to incorporate domain knowledge into their reward function design (Gupta et al. 2023). We examine the challenges of reward function design across three related schools of ethics, directly applicable to the kinds of domains AI systems are deployed into: disaster ethics, ecocentrism, and healthcare ethics.

Ethics

Disaster Ethics Disaster ethics is a branch of applied ethics, which involves the application of existing ethical schools to disaster scenarios. Given that disasters tend to affect large numbers of people, disaster consequentialist approaches, in which the consequences of an action are the focus, appear to fit the context of minimising deaths and suffering (Rakić 2018). In this paper, we focus on utilitarianism, in which the moral value of an action is determined by the utility, or good, brought about (Driver 2022).

Ecocentrism Researchers from across AI ethics and environmental philosophy have long critiqued approaches to AI ethics which centre on humans alone, advocating for broader environmental considerations (Rigley et al. 2023; Coeckelbergh 2022; Broo, Gellers, and Sætra 2024; Braidotti 2013). For this reason, we also include in our experiments ecocentrism, which extends moral consideration to entire ecosystems and to all active individuals within this. Ecocentrism posits that "a thing is right when it preserves the integrity of the biotic community, it is wrong when it tends otherwise" (Leopold 1949, p. 200).

Healthcare Ethics Healthcare ethics is another branch of applied ethics, in which established principles and theories are applied to healthcare specific cases. The most commonly used principles in healthcare ethics are non-maleficence (do no harm), beneficence (benefit only), autonomy (of the patient), and justice (treating all equally) (Summers 2009).

Proposed Method: Modelled-Ethics Value Alignment

Across many domains, humans do not make ethical decisions based solely on their own intuitions, but make use of evidence, advancements in data availability, and mathematical modelling. For instance, ecologists have long adopted mathematical models to more accurately capture and understand how ecosystems function, iteratively making use of big data and computational power to improve accuracy (Jørgensen and Fath 2011). Where an ethically concerning decision needs to be made - e.g., whether to cull an invasive species - ecologists make use of models to make the most accurate, and ethical, decision. So, when domain specialists use mathematical models to make ethically concerning decisions, instead of deferring to their own intuitions, it would seem intuitive that AI charged with making the same decisions should also be trained using those action guiding models.

Mathematical models have recently been introduced as an alternative to case based reward functions in the domain of autonomous vehicles (Abouelazm, Michel, and Zoellner 2024). For example, Nvidia Safety Force Field (SFF) models vehicle trajectories to minimise the intersection between an autonomous vehicle and other road users (Nistér et al. 2019). To the best of our knowledge, the use of mathematical models within RL reward function design remains a fringe approach compared to the more common hand-set, conditional reward functions, and has not been applied to the case of value alignment.

We propose an alternative approach to value alignment we call, 'modelled-ethics' (ME) value alignment. Put simply, we suggest that when an AI system is deployed into a specific domain, it should not be trained to align to the intuitions of the humans around it, but to the models which guide those humans in their decision processes. In some sense, we are cutting out the error-prone middle human.

To clarify, the approach put forward in this work does not replace the process of moral philosophy nor claim that normative ethics is detached from humans. We instead highlight the differences between hand-setting real number values and using nuanced mathematical models, in capturing and translating normative values. In this way, it is not the normative principles which come from the "the error-prone middle human", but the real number values used capture and translate those principles.

We define our ME approach to reward function design as R^m,

$$R^m(s,a,s') = \begin{cases} c_1 & s' = \text{terminal state}, \\ \mu(s') & \text{else} \end{cases} \quad (3)$$

where existing, established, domain-specific mathematical models, μ, calculate, for a given contextualised scenario, the moral value of being in some state, s, taking some morally concerning action, a, to end up in state s'. In our experiments, μ returns the value for all transitions except to the terminal state, which requires a value, c_1, to motivate the agent towards a goal. The expected value of state-action pairs is updated continuously as the agent explores, converging towards their true value. Through Q-learning, employing equation 1, the agent approximates the optimal action-value function Q* and pursues the optimal policy, i.e., the policy which maximises ethical values modelled by μ.

Comparing our ME approach in Equation 3 to the hand-setting approach in Equation 2, we side-step the need to: a) explicate all moral scenarios and b) manually tune all corresponding values.

Experimental Design

With this paper, we compare how well the policies output by hand-programmed and ME agents align with ethical principles in three related use case scenarios, directly relevant to our three schools of ethics (disaster, ecology, and health). We abstracted each use case as a 20x20 gridworld. We based our gridworld environment and Q-learning algorithm code on that of (Tinsley 2018), heavily adapted to fit our case study and reconfigured for various comparative experiments. To improve the reliability of our results, we built, for each ethical use case, 5 gridworlds (GWs 1-5) with random start and terminal states[1].

In the disaster ethics case, a UAV chooses a flight path for dropping emergency aid within a disaster-stricken area. Each GW contains 5 randomly allocated 'hotspots', with higher GDPs and populations, and therefore lives affected and risks of death. 5 gave the agent a challenging environment, where they may try to reach multiple hotspots, whilst not clogging the small 20x20 grid space with only high value states. In the ecological case, a UAV chooses a flight path through a game reserve at the risk of scaring away wildlife. We used Bennitt et al. (2019)'s open access data on the reactions of terrestrial mammals to UAVs. The populations of wildlife, locations, and responses to drones were unique for the 5 GWs. In the healthcare case, a UAV delivers a defibrillator to an isolated cardiac arrest victim, where delays of even seconds can cause serious, even fatal, harms to the patient. For the healthcare cases, we set a drone speed of 12.5m/s, based on the drone speeds recorded in (Bennitt et al. 2019) and accounting for the slower speed as a result of carrying the defibrillator.

To hand-program reward functions, we iteratively set rewards, ran experiments, and tuned the values to improve the performance of the agent. This is standard practice, as self-reported by RL researchers (Knox et al. 2023; Booth et al. 2023). We repeated this process 5 times per use case scenario, one more than the average number of iterations recorded by Knox et al. (2023). We tuned our hyperparameters (learning rate, α, exploration rate, ϵ, discount factor, γ, and maximum episodes) using a mixture of manual and automated techniques. For automated hyperparameter optimisation, we used the Optuna package[2], as a combination of independent sampling and Bayesian optimisation to converge toward the best set of hyperparameters (Akiba et al. 2019a,b; Baldé 2023). The disaster (hand-programmed and ME) hyperparameters are $\epsilon = 0.8$, $\alpha = 0.59$, and $\gamma = 0.96$; for ecology (hand-programmed and ME) and healthcare (ME),

[1]Random processes were seeded from 42 onwards

[2]Optuna v3.6.1 for Python v3.11.5

$\epsilon = 0.01$ $\alpha = 0.1$ and $\gamma = 0.9$; and for healthcare (hand-programmed), $\epsilon = 0.13$ $\alpha = 0.7$ and $\gamma = 0.6$. All algorithms were run for maximum of 35000 trials.

In total, we ran 90 experiments[3]. For each of our three ethical use cases, we built 5 GWs. For each GW, we hand-programmed 5 reward functions and compared them against our own modelled-ethics reward function.

Ethical Implementation

For fair comparison, the ME and hand-set reward functions of each ethical theory employ the same state features. For disaster ethics, both reward functions employ state GDP, area, and population; for ecocentrism, both functions employ wildlife population change; and for healthcare, both functions employ time from state to patient.

Disaster Ethics In line with our definition of disaster utilitarianism, the agent ought to select states which would benefit most from aid deliveries. We base our ME approach for the disaster scenario on the work of Song and Park (2019), who employ multiple regression and correlation analysis of data on populations, area sizes, and GDPs to present three equations predicting human lives lost (HLD), human lives affected (HLA) and damage costs (DCS) as a result of some disaster. These equations are defined in Equation 4 below,

$$\begin{gathered} HLD = 975.7635 - 0.4389x_1(s') + 0.0004x_2(s') + \\ 0.0702x_3(s'), \\ HLA = 205644.9682 - 96.7326x_1(s') + 0.0954x_2(s') + \\ 18.0191x_3(s'), \\ DCS = 17968.0283 + 476.6021x_1(s') + 0.0425x_2(s') + \\ 0.2442x_3(s') \end{gathered} \tag{4}$$

where $x_1(s')$, $x_2(s')$, and $x_3(s')$ refer to GDP, area, and population of state s', respectively. Because natural disasters affects national economic growth, population, and GDP, and death rate is significantly affected by the negative effects of the ecological economy and GDP (Song and Park 2019), we assume equal importance of HLD, HLA, and DCS. We multiplied these values together to provide a single value outputting the overall predicted damage as a result of a disaster, as a numerical estimation of need for aid. We shifted the outputs of the damage prediction formula to below zero, to avoid the agent circling around high value states. Our ME disaster reward function, R^m is defined below,

$$\begin{gathered} R^m(s,a,s') = \begin{cases} 50 & s' = \text{terminal state}, \\ \mu(s') & \text{else} \end{cases}, \\ \mu(s') = HLD(s') \cdot HLA(s') \cdot DCS(s') - \\ \max_{\overline{s} \in S} (HLD(\overline{s}) \cdot HLA(\overline{s}) \cdot DCS(\overline{s})) \end{gathered} \tag{5}$$

where the reward for being in state s and taking action a to end up in state s' is calculated as the product of human lives lost (HLD), lives affected (HLA), and damage costs (DCS) incurred in state s', shifted below zero (Song and Park 2019). The agent is also rewarded $+50$ for reaching the terminal state. Figure 1a plots an example gridworld in which the predicted damage has been calculated for all states using equation 5. The start and terminal states are white stars and the policy output by the ME agent is the white line.

An intuitive approach to hand-programming a reward function (R^h) for the disaster UAV is to reward the agent for selecting a route through hotspots, wherein the values for x_1 and x_3 are highest - since each state has the same area, x_2 is equal across all states. This abstracted below.

$$R^h(s,a,s') = \begin{cases} c_1 & s' = \text{`hotspot'}, \\ c_2 & s' = \text{terminal state}, \\ c_3 & \text{else} \end{cases} \tag{6}$$

Across the 5 iterations of this reward function design, we tried combinations of $c_1 = [10, 50, 100]$, $c_2 = [0, 50, 100]$, and $c_3 = -1$. We found that $c_1 = 10$ and $c_2 = 100$ led to the best performance in this scenario. Figure 1d shows the first iteration of our hard-programmed reward function design applied to GW3, in which the hotspots are highlighted in yellow, the start and terminal states as white stars, and the policy chosen by the agent as a white line. In this iteration, $c_1 = 50$, $c_2 = 0$, and $c_3 = -1$.

Ecocentrism In accordance with the principles of ecocentrism, the UAV should select a path which avoids disturbing wildlife and, thereby, the wider ecosystem. We define our ME ecocentric reward function as calculating the root mean squared error between the ecosystem as it is and as a result of the actions of an AI system (Equation 8). Since changes to the dynamics between species in turn change the wider ecosystem, we model the ecosystem using the Lotka-Volterra model of predator-prey dynamics or, more broadly, dynamics between two or more species (Wangersky 1978). This is "one of the best examples of simple tractable models in ecology" (AlAdwani and Saavedra 2020, p. 1). The equations for a density dependent Lotka-Volterra model of predator and prey can be expressed as,

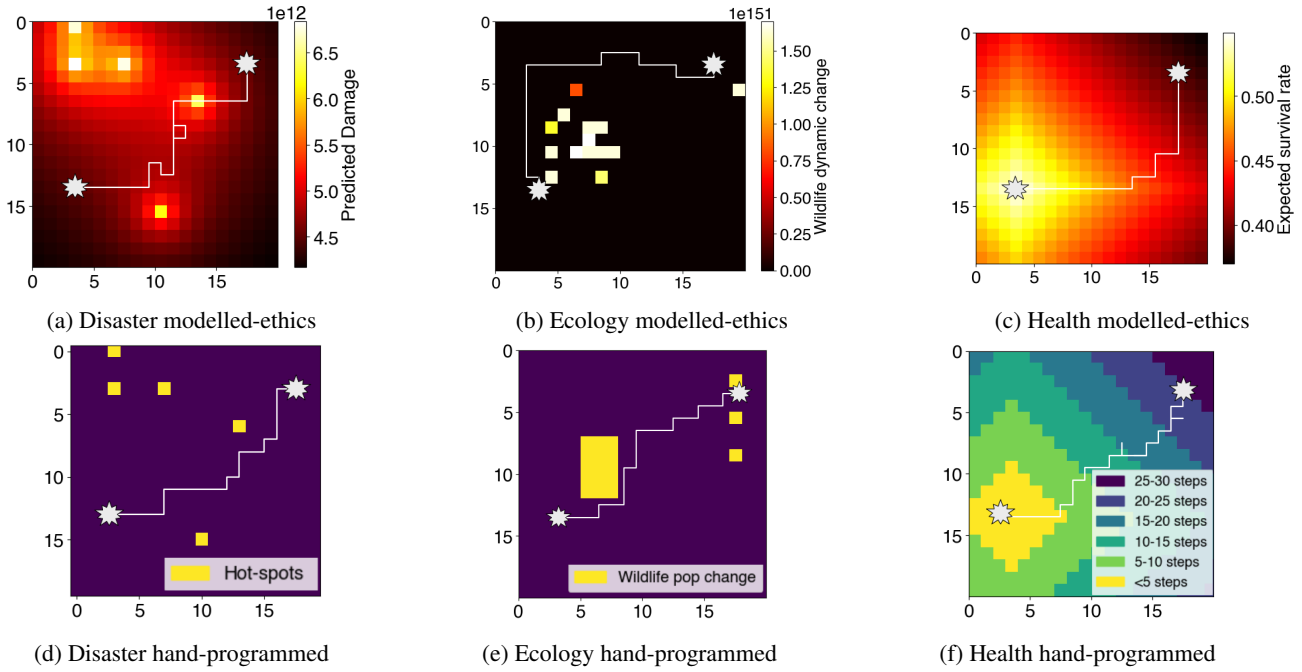
$$\begin{aligned} \frac{dx}{dt} &= rx[1 - \frac{x}{K}] - axy, \\ \frac{dy}{dt} &= \beta xy - Dy. \end{aligned} \tag{7}$$

where x represents the number of prey organisms, r the growth rate, y the number of predator organisms, a a proportionality constant linking the prey mortality to the number of prey and predators, K the carrying capacity of the environment in terms of species population, β a proportionality constant linking the increase in predators to the number of prey and predators, and D a constant of mortality for the predators (Wangersky 1978, eq. 11).

With this model of the ecosystem, our ME ecocentric reward function is defined as:

$$\begin{gathered} R^m(s,a,s') = \begin{cases} 50 & s' = \text{terminal state}, \\ \mu(s') & \text{else} \end{cases}, \\ \mu(s') = -\sqrt{(LV_{ideal}(s') - LV_{real}(s'))^2} \end{gathered} \tag{8}$$

[3]All experiments were run on a 2.6 GHz 6-Core Intel Core i7 processor



(a) Disaster modelled-ethics (b) Ecology modelled-ethics (c) Health modelled-ethics

(d) Disaster hand-programmed (e) Ecology hand-programmed (f) Health hand-programmed

Figure 1: Reward functions and policies for Gridworld 2, where stars represent start and terminal states and lines represent policies.

where $LV_{real}(s')$ provides a Lotka-Volterra model of the ecosystem at state s', where populations are affected by the presence of the UAV, and $LV_{ideal}(s')$ provides a model of the ideal version of the ecosystem at state, s', where the UAV has not disturbed any wildlife. This is negated since $R^m(s, a, s')$ is a function of environmental disturbance our agent ought to minimise. If the change is negated, the agent will pursue actions with a higher value and which therefore cause less change. Our Lotka Volterra code is adapted from (Magnani 2021) to include population changes and model comparisons for the proposed reward function. Figure 1b shows an example gridworld in which this model has been applied.

An intuitive hard-programmed reward function (R^h) for the ecological UAV would apply punishments for every state-action pair in which an animal is scared away from the area. This has been abstracted below,

$$R^h(s, a, s') = \begin{cases} c_1 & \text{wildlife population change,} \\ c_2 & s' = \text{terminal state,} \\ c_3 & \text{else} \end{cases} \tag{9}$$

In the 5 iterations of this hand-programmed reward function design, we tried a combination of $c_1 = [-10, -50]$, $c_2 = [10, 50, 100]$, and $c_3 = [-1, 0]$. Overall, we found that a harsh punishment $[-50]$ for c_1 and low value $[10]$ for c_2 led to the best performance in this scenario. Figure 1e shows the outputs of the first iteration of our hard-programmed reward function design in GW3, where yellow states are those in which wildlife would be scared away should the UAV enter. Here, $c_1 = -10$, $c_2 = 100$, and $c_3 = 0$.

Healthcare Ethics In accordance with healthcare ethics, the agent ought to avoid delays of the medicine (non-maleficence) and maximise the expected survival of the cardiac-arrest patient (beneficence). Our ME approach applies a power-law function which "best reflects the behavior of the survival rate [of the cardiac arrest patient] as a function of time t to defibrillation" (Pourghaderi et al. 2022, p. 216). Applying this model to the reward function ensures the agent will be incentivised to pursue states with a higher expected survival rate (*ESR*) of the patient. The ME reward function for the healthcare agent is defined below,

$$R^m(s, a, s') = \begin{cases} 50 & s' = \text{terminal state,} \\ \mu(s') & \text{else} \end{cases},$$
$$\mu(s') = ESR(t_{s'}) - \max_{\bar{s} \in S} ESR(t_{\bar{s}}), \tag{10}$$
$$ESR(t_{s'}) = 0.549{t_{s'}}^{-0.584}$$

where the reward for a given state-action pair, s, a which leads to state s' corresponds to the expected survival rate of the patient given the time taken in minutes, $t_{s'}$, to get from state s' to the patient. This is shifted below zero by subtracting the maximum expected survival rate across the state space, to avoid the agent circling around high value states. An extra minute is also added to account for the time taken to pack and send off the UAV, and to set up the defibrillator.

Figure 1c shows an example of this ME reward function applied in GW3. We used the same Manhattan distance function for both the ME and hand-programmed reward functions for this healthcare case, where the total distance from one state to another is the sum of horizontal and vertical dis-

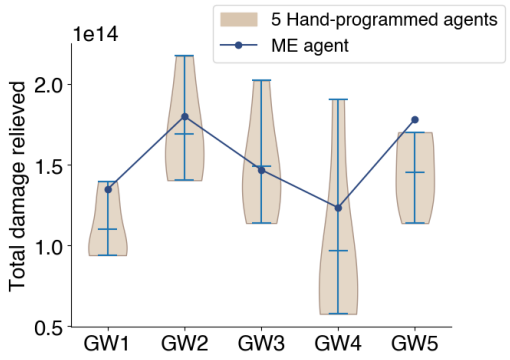


Figure 2: Total damage relieved across policy.

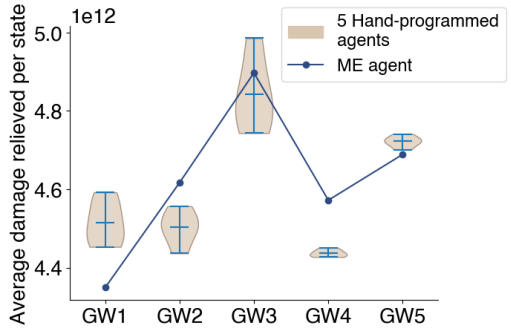


Figure 3: Average damage relieved per state across policy.

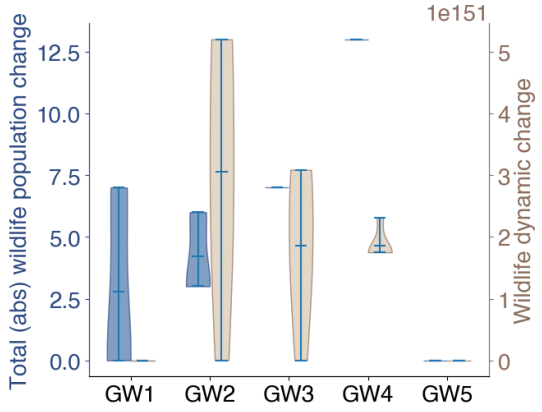


Figure 4: Changes to wildlife populations and to dynamics caused by hand-programmed agent's policies.

tances, given that the agent can only move horizontally or vertically.

An intuitive hand-programmed reward function would apply costs to delays. For the 5 iterations of hand-programmed reward design, we applied an accumulating cost, c_1, for delays of various times, t, and reward for reaching the terminal state, c_2. We tried combinations of $c_1 = [-1, -10]$, $t = [10, 20 \text{ seconds}]$, and $c_2 = [0, 50]$. We also tried a punishment of -1 for every state spent, which was the most effective reward function design. Figure 1f shows the first iteration of our hand-programmed reward function design for GW3 in which the distance, and therefore time taken, from a given state to the terminal state are plotted in 5-block intervals. Here, $c_1 = 10$, $t = 10$ seconds, and $c_2 = 50$.

Experimental Limitations

We had a single programmer work through the iterative process of hand setting reward functions, which may have introduced bias or subjectivity. Part of the motivation for this paper is the methodological limitations of relying on programmers to hand-set reward functions due to this bias and subjectivity. Nevertheless, we aimed to ensure the methodology for hand-programmed design was replicable and matched the standard approach as identified by the literature (Booth et al. 2023; Knox et al. 2023).

Results and Discussion

Disaster Ethics

To examine alignment to disaster utilitarianism, we calculated the total (across all states) and average (for each state) predicted damage within the agent's policy to measure the overall damage relieved and the agent's focus on high-risk areas. We assume that an agent aligned with utilitarian disaster ethics would choose a route which maximises total and average damage relieved.

Figure 2 shows the total damage relieved in the policies chosen by the 5 hand-programmed agents (violins) and single ME agent (line). Within each gridworld, we can see a high variance among the outputs of the 5 hand-programmed agents, especially in GW4. Figure 3 shows the average damage relieved across all states in the policies chosen by the hand-programmed agents (violins) compared to the ME agent (line). Here, there is less variance among the 5 hand-programmed reward functions, though still substantial variance in GW3. This shows that the results of the hand-programmed agent depend heavily on the values input by the programmer.

Figure 2 also shows that in 4/5 gridworlds, the ME agent's policies result in higher total damage relieved than the average results across 5 iterations of hand-programmed reward function design (middle line of the violin). Figure 3 shows that in 3/5 gridworlds, the ME agent's policies also result in higher average damage relieved compared to the average results of the 5 hand-programmed reward function designs. However, in both Figures 2 and 3, we can see that, in some GWs, the maximum results achieved of the 5 hand-programmed agents (top line of the violin) is higher than the ME agent's results. Together, these results indicate that some hand-programmed agents outperformed our ME agent, but on average, the hand-programmed agents performed worse.

Ecocentrism

We measured changes to wildlife dynamics and to wildlife population as a result of the agent's policies. We assume an agent aligned with ecocentrism would minimise these values. Figure 4 plots the results for the hand-programmed agents. Across all 5 gridworlds, the ME agent's policies led

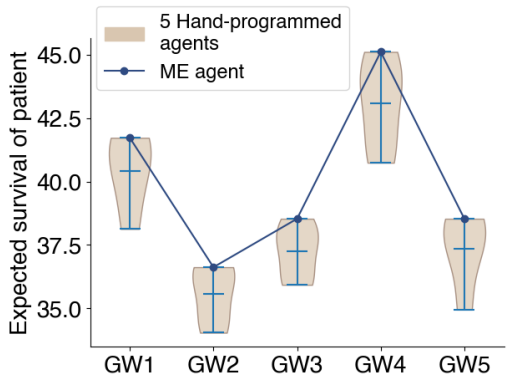


Figure 5: Expected survival rate of patient.

to no wildlife dynamic change nor population change; resulting in 0 for both metrics across all 5 gridworlds. We didn't plot these ME results.

As Figure 4 clearly shows, the hand-programmed agent's policies led to significantly higher total wildlife population and dynamic changes than the ME agent. We found a statistically significant difference [Kruskal-Wallis p-value $<$ 0.05] in both the population and dynamic results between the hand-programmed and ME agents. There is also, again, variance in the efficacy of the hand-programmed agent in some environments, for one or both metrics; in GW2 there is variance across both metrics, in GWs 1, 3, and 5 there is variance in one metric, and consistency in the other. This shows the results of the hand-programmed agent again depend on the values chosen by the programmer and the environment.

Healthcare Ethics

We measured the expected survival rate of the patient, plotted in Figure 5. We assume that an agent which acts in accordance with healthcare ethics would maximise expected survival rate.

Overall, the results of the ME agent's policies are better (higher expected survival) than the average results of the hand-programmed agents. However, the ME agent never completely outperformed the best hand-programmed agent, as results for the ME agent equal the highest results for the hand-programmed agents. This is perhaps because in simple tasks (e.g., getting from start to terminal in as few steps as possible), a simple reward function (e.g., punishing every state spent) will suffice.

Discussion

Across all scenarios and most GWs, there is variance in the results of the hand-programmed agent, showing dependency on the real number values chosen by the programmer. Though some iterations of our hand-programmed reward function design could outperform our proposed ME, this was not consistent and was highly dependent on the ethical scenario, environment, and values assigned. Our ME approach mostly out performed our hand-programmed agents.

Ethicists may object to our approach, arguing that ethical principles are not domain-specific (Kaas 2021). However, even if moral duties are consistent across domains, their practical execution requires information specific to the case at hand. At a high level of abstraction, one could view all of our RL agents as maximising utility. But at ground level, the executions are very different.

In Figure 1, we can see that in the disaster and healthcare cases, the ME reward functions include more granular information about the value of moving between states. In the ecological case, this information was not only more granular, but different; the positioning of the lighter states with higher disturbance calculated by the ecological model (Figure 1b) differ from the yellow states in which wildlife population would change should the agent enter (Figure 1e). The difference between the hand-programmed and ME agents in terms of aligning to moral values was also most stark in the ecological case as the hand-programmed agent oftentimes disturbed wildlife, despite the costs of doing so built into that system's architecture. This shows that mathematical models can better align with ethical theories than human intuitions, particularly when domain-specific knowledge is crucial.

One drawback to our domain-specific reward functions is their limited generalization to diverse scenarios (Abouelazm, Michel, and Zoellner 2024). Our approach requires careful, informed model selection and setting the value of the terminal state. For instance, in equation 8, though the models for LV_{ideal} and LV_{real} are pre-defined, one may still have to optimise the cost function between these two models. For this reason, there may be certain cases and environments where using a hand-programmed reward function would be easier and just as effective - such as in our healthcare scenario - or even more effective - such as in our disaster scenario. However, our results show that hand-setting values is an unreliable approach, depending heavily on the choices of an individual programmer, environment, and scenario. In high-stakes cases, value alignment requires reliable, consistent methods, and it is important that ethical risks are not left solely for individual programmers to avoid.

Conclusion

This paper examines the challenge of hand-programming ethical values into reward functions for aligned RL systems. We proposed an alternative approach based on mathematically modelling ethical values. Our results indicate that the efficacy of hand-programmed reward functions depend heavily on the ethical scenario, environment, and numeric values assigned in reward design. Our ME approach offers a more consistent alternative and outperformed our hand-programmed reward functions on average. An important direction for future research is to expand the complexity of environments to that of real-world applications, considering how the efficacy of different domain-specific models will alter across environments and case studies.

Acknowledgments

The authors would like to acknowledge support from the Defence and Security Programme at the Alan Turing Institute, funded by the UK Government.

References

Abel, D.; MacGlashan, J.; and Littman, M. L. 2016. Reinforcement Learning as a Framework for Ethical Decision Making. In Bonet, B.; Koenig, S.; Kuipers, B.; Nourbakhsh, I. R.; Russell, S.; Vardi, M. Y.; and Walsh, T., eds., *AI, Ethics, and Society, Papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 13, 2016*, volume WS-16-02 of *AAAI Technical Report*. AAAI Press.

Abouelazm, A.; Michel, J.; and Zoellner, J. M. 2024. A Review of Reward Functions for Reinforcement Learning in the context of Autonomous Driving. *arXiv*.

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019a. Optuna. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM.

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019b. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

AlAdwani, M.; and Saavedra, S. 2020. Ecological models: higher complexity in, higher feasibility out. *Journal of The Royal Society Interface*, 17(172): 20200607.

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

Baldé, B. 2023. Bayesian Sorcery for Hyperparameter Optimization using Optuna. *Medium*, (8 June). [Accessed 31 July 2024].

Bennitt, E.; Bartlam-Brooks, H. L. A.; Hubel, T. Y.; and Wilson, A. M. 2019. Terrestrial mammalian wildlife responses to Unmanned Aerial Systems approaches. *Scientific Reports*, 9(1).

Booth, S.; Knox, W. B.; Shah, J.; Niekum, S.; Stone, P.; and Allievi, A. 2023. The Perils of Trial-and-Error Reward Design: Misdesign through Overfitting and Invalid Task Specifications. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5): 5920–5929.

Braidotti, R. 2013. *The Posthuman*. Polity Press.

Broo, D. G.; Gellers, J. C.; and Sætra, H. S. 2024. Re-imagining Intelligent Machines in an Anthropocentric–Ecocentric Continuum: The Case for Ecocentric Intelligent Machines. *Journal of Industrial Information Integration*, 100636.

Brundage, M. 2014. Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3): 355–372.

Coeckelbergh, M. 2022. *Robot Ethics*. MIT Press.

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

Ecoffet, A.; and Joel, L. 2021. *Reinforcement Learning Under Moral Uncertainty*.

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

Gupta, D.; Chandak, Y.; Jordan, S. M.; Thomas, P. S.; and da Silva, B. C. 2023. Behavior Alignment via Reward Function Optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hibbard, B. 2014. Ethical artificial intelligence. *arXiv preprint arXiv:1411.1373*.

Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Zhou, J.; Zhang, Z.; Zeng, F.; Ng, K. Y.; Dai, J.; Pan, X.; O'Gara, A.; Lei, Y.; Xu, H.; Tse, B.; Fu, J.; McAleer, S.; Yang, Y.; Wang, Y.; Zhu, S.-C.; Guo, Y.; and Gao, W. 2024. AI Alignment: A Comprehensive Survey. arXiv:2310.19852.

Jørgensen, S. E.; and Fath, B. D. 2011. 1 - Introduction. In Jørgensen, S. E.; and Fath, B. D., eds., *Fundamentals of Ecological Modelling*, volume 23 of *Developments in Environmental Modelling*, 1–18. Elsevier.

Kaas, M. H. 2021. Raising Ethical Machines: Bottom-Up Methods to Implementing Machine Ethics. In *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*, 47–68. IGI Global.

Knox, W. B.; Allievi, A.; Banzhaf, H.; Schmitt, F.; and Stone, P. 2023. Reward (Mis)design for autonomous driving. *Artificial Intelligence*, 316: 103829.

Leopold, A. 1949. The Land Ethic. In Keller, D. R., ed., *Environmental Ethics: The Big Questions*, 193–201. Chichester: Wiley-Blackwell.

Magnani, F. 2021. Lotka Volterra N Species Model. https://github.com/FMagnani/Generalized-Lotka-VolterraN-species-model. [Accessed 7 September 2023].

Metz, F.; and Bukov, M. 2023. Self-correcting quantum many-body control using reinforcement learning with tensor networks. *Nature Machine Intelligence*, 5(7): 780–791.

Ng, A. Y.; and Russell, S. J. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, 663–670. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Nistér, D.; Lee, H.-L.; Ng, J.; and Wang, Y. 2019. The Safety Force Field. Technical report, Nvidia.

Noothigattu, R.; Gaikwad, S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. 2018. A Voting-Based System for Ethical Decision Making. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Peschl, M. 2021. Training for Implicit Norms in Deep Reinforcement Learning Agents through Adversarial Multi-Objective Reward Optimization. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 275–276. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Pourghaderi, A. R.; Kogtikov, N.; Lees, M. H.; Cai, W.; Pin Pek, P.; Fu Wah Ho, A.; Ming Ng, W.; Kwak, J.; Elgin White, A.; Lynn Lim, S.; Shao Wei Lam, S.; and Eng Hock Ong, M. 2022. Maximum expected survival rate model for public access defibrillator placement. *Resuscitation*, 170: 213–221.

Rakić, V. 2018. Disaster Consequentialism. In *Advancing Global Bioethics*, 145–156. Springer International Publishing. ISBN 9783319927213.

Rigley, E.; Chapman, A.; Evers, C.; and McNeill, W. 2023. Anthropocentrism and Environmental Wellbeing in AI Ethics Standards: A Scoping Review and Discussion. *AI*, 4(4): 844–874.

Rodriguez-Soto, M.; Rădulescu, R.; Rodriguez-Aguilar, J. A.; Lopez-Sanchez, M.; and Nowé, A. 2023. Multi-objective reinforcement learning for guaranteeing alignment with multiple values. In *Proc. of the Adaptive and Learning Agents Workshop (ALA 2023), Cruz, Hayes, Wang, Yates (eds.)*.

Rodriguez-Soto, M.; Serramia, M.; Lopez-Sanchez, M.; and Rodriguez-Aguilar, J. A. 2022. Instilling moral value alignment by means of multi-objective reinforcement learning. *Ethics and Information Technology*, 24(1): 9.

Song, Y. S.; and Park, M. J. 2019. Development of Damage Prediction Formula for Natural Disasters Considering Economic Indicators. *Sustainability*, 11(3): 868.

Summers, J. 2009. Principles of Healthcare Ethics. In Morrison, E. E., ed., *Health Care Ethics: Critical Issues for the 21st Century*, chapter 2. Jones and Bartlett Publishers, 2nd edition.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning : an introduction*. Cambridge, Massachusetts: The MIT Press, second edition edition.

Tinsley, M. 2018. Gridworld-with-Q-Learning-Reinforcement-Learning-. https://github.com/michaeltinsley/Gridworld-with-Q-Learning-Reinforcement-Learning-. [Accessed 17 November].

Tolmeijer, S.; Kneer, M.; Sarasua, C.; Christen, M.; and Bernstein, A. 2020. Implementations in Machine Ethics: A Survey. *ACM Computing Surveys*, 53: 1–38.

Trussell, H. J. 2018. Why a Special Issue on Machine Ethics. *Proceedings of the IEEE*, 106(10): 1774–1776.

Vishwanath, A.; Dennis, L. A.; and Slavkovik, M. 2024. Reinforcement Learning and Machine ethics:a systematic review. arXiv:2407.02425.

Wangersky, P. J. 1978. Lotka-Volterra Population Models. *Annual Review of Ecology and Systematics*, 9(1): 189–218.

Watkins, C. J. C. H.; and Dayan, P. 1992. Q-learning. *Machine Learning*, 8(3-4): 279–292.

Wu, Y.-H.; and Lin, S.-D. 2017. A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.