Operationalising Rawlsian Ethics for Fairness in Norm-Learning Agents

Jessica Woodgate, Paul Marshall, Nirav Ajmeri

School of Computer Science
University of Bristol
Bristol BS8 1UB, UK
jessica.woodgate@bristol.ac.uk, p.marshall@bristol.ac.uk, nirav.ajmeri@bristol.ac.uk

Abstract

Social norms are standards of behaviour common in a society. However, when agents make decisions without considering how others are impacted, norms can emerge that lead to the subjugation of certain agents. We present RAWL·E, a method to create ethical norm-learning agents. RAWL·E agents operationalise maximin, a fairness principle from Rawlsian ethics, in their decision-making processes to promote ethical norms by balancing societal well-being with individual goals. We evaluate RAWL·E agents in simulated harvesting scenarios. We find that norms emerging in RAWL·E agent societies enhance social welfare, fairness, and robustness, and yield higher minimum experience compared to those that emerge in agent societies that do not implement Rawlsian ethics.

1 Introduction

Social norms are standards of expected behaviour that govern a multi-agent system (MAS) and enable coordination between agents (Levy and Griffiths 2021; Wright 1963). Norms can be established through top-down prescriptions or emerge bottom-up via interactions between agents (Morris-Martin, De Vos, and Padget 2019). However, when agents are solely self-interested, norms may emerge that exploit some agents for the benefit of others. Where ethics involves one agent's concern for another (Murukannaiah and Singh 2020), norms which result in the subjugation of agents are unethical. If agents learn norms by appealing to existing behaviours in a society without evaluating how ethical those behaviours are, they risk perpetuating unethical norms.

Previous works on promoting norms which are considerate of others, such as Tzeng et al. (2022) and Dell'Anna et al. (2020), appeal to individual or societal preferences over values. Other works observe the behaviour of others to encourage cooperation: Oldenburg and Zhi (2024) infer norms by observing apparent violations of self-interest; Guo et al. (2020) learn contextual priority of norms from observing experts; Chen et al. (2017) imply norms through reciprocity.

However, approaches which appeal to preferences or existing behaviours to promote cooperation define ethical behaviour by reference to descriptive statements, which are statements that express what states of affairs are like (Kim, Hooker, and Donaldson 2021). Attributing ethics descriptively may lead to the issue of deriving an *ought* from an *is*—just because something is the case doesn't mean it ought to be. Where existing norms or behaviours are unethical, approaches that encourage cooperation through descriptive facts thereby risk propagating unethical norms which reflect *what is the case*, rather than *what ought to be*.

To mitigate the *is-ought* gap, we turn to *normative ethics*. Normative ethics is the study of practical means to determine the ethical acceptability of different courses of action (Woodgate and Ajmeri 2024). Normative ethics principles are justified by reason in philosophical theory. These principles are normative in that they are prescriptive, indicating how things ought to be, rather than descriptive, indicating how things are (Kim, Hooker, and Donaldson 2021).

The principle of *maximin*—to maximise the minimum experience—is a widely respected fairness principle in normative ethics advanced by Rawls (1958). Rawls states that in a society with unequal distribution which is not to the benefit of all, those benefiting the least should be prioritised. We hypothesise that creating agents that promote the emergence of ethical norms, while avoiding the *is-ought* gap, is aided by an appeal to Rawlsian ethics (Woodgate and Ajmeri 2022).

Contribution We propose RAWL·E, a novel method to design socially intelligent norm-learning agents that consider others in individual decision making by operationalising the principle of maximin from Rawlsian ethics. A RAWL·E agent includes an ethics module that applies maximin to assess the effects of its behaviour on others.

Novelty Operationalising Rawlsian ethics in learning agents to enable explicit norm emergence is a novel contribution. RAWL·E goes beyond existing works on norm-learning agents: Ajmeri et al.'s (2020) agents incorporate ethical decision-making, but do not incorporate learning. Agrawal et al. (2022) address the emergence of explicit norms, but optimise norms based on the sum of payoffs received by other agents, which might be unfair for some agents. Zimmer et al. (2021) and Balakrishnan et al. (2022) operationalise Rawlsian ethics in learning agents, but do not consider the role of norms. As a RAWL·E agent gains experience, it learns to achieve its goals whilst behaving in ways that support norms which prioritise those who are least advantaged in situations of unequal resource distribution.

We evaluate RAWL·E agents in two simulated harvesting scenarios implemented in reinforcement learning (RL). We find that (1) RAWL·E agents learn ethical norms that promote the well-being of the least advantaged agents, and (2) RAWL·E agent societies yield higher social welfare, fairness, and robustness than agents who do not operationalise Rawlsian ethics in their individual decision-making.

Organisation Section 2 explores related works and gaps. Section 3 describes our method. Section 4 presents the simulation environment used to evaluate RAWL·E agents. Section 5 discusses results of our simulation experiments. Section 6 concludes with a discussion of future directions.

2 Related Works

Research on combining normative ethics with norm emergence and learning is relevant to our contributions.

Normative Ethics and Norm Emergence Ethical norm emergence has been examined through the lens of agent roles. Anavankot et al. (2023) propose norm entrepreneurs that influence the dynamics of norm-following behaviours and thus the emergence of norms. Vinitsky et al. (2023) study norm emergence through sanction classification. Levy and Griffiths (2021) manipulate rewards using a central controller to enable norm emergence. Neufeld et al. (2022) use deontic logic to implement a normative supervisor module in RL agents. Yaman et al.'s (2023) agents sanction one another to encourage effective divisions of labour. Maranhão et al. (2022) formally reason about normative change. However, a gap remains in agents learning norms based on what ought to be the case, rather than what is. We address this gap by implementing principles from normative ethics to encourage the emergence of norms that can be justified independently to a specific situation.

Traditional approaches encourage norm emergence by maximising social welfare—how much society as a whole gains. Shoham and Tennenholtz (1997) promote highest cumulative reward. Yu et al. (2014) utilise majority vote. Agrawal et al. (2022) sum the payoffs for different stakeholders. Focusing on social welfare alone may lead to situations where a minority is treated unfairly for the greater good (Anderson, Anderson, and Armen 2004), and mutual reward does not specify how to coordinate fairly (Grupen, Selman, and Lee 2022). To mitigate weaknesses associated with only maximising social welfare, we implement Rawlsian ethics, emphasising improving the minimum experience.

Normative Ethics and Learning Jing and Doorn (2020) emphasise the importance of focusing on positive standards alongside preventative ethics, which involves negative rules denoting wrongdoing. As ethics is dynamic, it may not always be possible to determine which behaviours to restrict.

Svegliato et al. (2021) implement divine command theory, prima facie duties, and virtue ethics; Nashed et al. (2021) implement the veil of ignorance, golden rule and act utilitarianism. Dong et al. (2024) optimise federated policies under utilitarian and egalitarian criteria. A gap exists, however, in applying normative ethics in RL to norm emergence. RAWL·E addresses that gap.

3 Method

We now present our method to design RAWL·E agents who operationalise Rawlsian ethics to support the emergence of ethical norms.

3.1 Schematic

Definition 1. *Environment $\mathbb{E}$ is a tuple $\langle AG, D, \mathcal{N} \rangle$ where, $AG = \{ag_1, ..., ag_n\}$ is a set of agents; D is the amount of total resources; $\mathcal{N}$ is the set of norms.*

Definition 2. *A RAWL·E agent is a tuple : $\langle d, v, G, A, Z, \mathsf{NM}, \mathsf{EM}, \mathsf{DM} \rangle$ where, $d \in D$ is the amount of resources to which the agent has access; v is a measure of its well-being; G is the set of goals $g_1, ..., g_l$; A are the actions available to the agent to help achieve its goals; Z are the behaviours which the agent has learned;* NM *is its norms module;* EM *is its ethics module; and* DM *is its interaction module.*

Definition 3. *A goal $g \in G$ is a set of favourable states an agent aims to achieve.*

Definition 4. *A behaviour $\zeta \in Z$ is a tuple $\langle \mathsf{pre}, \mathsf{act} \rangle$, where* pre $\in$ *Expr is its precondition;* act $\in$ *Expr is its action; and Expr is any logical expression that can be evaluated as either true or false based on the values of its variables.*

A behaviour has a precondition denoting the conditions within which the behaviour arises, and a postcondition, which is the action implied by the precondition. Each agent keeps a record of their learnt behaviours.

A behaviour is encoded in the form of an if-then rule:

```
<behaviour> ::= IF <pre> THEN <act>
```

Definition 5. *A norm $n \in \mathcal{N}$, where $\mathcal{N} \subseteq Z$, is a behaviour adopted by a society.*

Norms are the prescription and proscription of agent behaviour on a societal level (Savarimuthu et al. 2013).

Definition 6. *$\mathcal{N}$, where $\mathcal{N} \subseteq Z$, denotes the set of emerged norms, i.e., the behaviours adopted by the society as norms, which form a normative system describing a society.*

Norms emerge when the same behaviours are adopted by other agents (Tuomela 1995). Norm emergence is accepted to have happened when a predetermined percentage of the population adopt the same behaviours. As following previous literature, we assume a norm to have emerged when it reaches 90% convergence (Kittock 1995).

Definition 7. *A sanction F represents a positive or negative reaction to behaviour which provides feedback to the learner in the form of a reward.*

Sanctions are positive or negative reactions to behaviour which help enforce norms. A self-directed sanction is a sanction directed towards and affecting only its sender (Nardin et al. 2016). The self-directed sanction provides feedback to the learner as a reward.

3.2 Interaction and Norm Learning

To make decisions and pursue their goals, RAWL·E agents use ethics module, norms module, and interaction module.

Ethics Module Ethics module, EM, assesses how actions affect the well-being of other agents. To evaluate the well-being of others, RAWL·E agents implement Rawlsian ethics. Adapted from Leben (2020), an ethical utility function $u(d) \rightarrow (\upsilon)$ models a distribution of resources, where d is a vector of resource distribution which sums to D, the amount of total resources, and (υ) is a measurement of well-being for agents considering that resource distribution. Where w is a vector of inputs (e.g., observed well-being of agents), Rawlsian ethics is expressed as:

$$MA(d) = min_w \quad u(d, \upsilon_i) \tag{1}$$

Via $MA(d)$, the ethics module evaluates whether the agent's action improves the minimum experience. It generates a positive self-directed sanction ξ if an action improves the minimum experience, and a neutral or negative sanction $-\xi$ if it does not change or worsens. In analogy to the real world, a positive sanction represents happiness from helping others, while a negative sanction represents guilt. To implement MA, ethics module takes as input U_t and U_{t+1}, where U is a vector of well-being $\upsilon_1, \ldots, \upsilon_n$ for all agents $ag_1, \ldots, ag_n$ at times t and $t+1$. Ethics module identifies the minimum experience $min_w u(d, \upsilon)$ at t and $t+1$, storing the results in $\upsilon_{\min t}$ and $\upsilon_{\min t+1}$, respectively. Therefore:

$$F_{t+1}(s_t, s_{t+1}) = \begin{cases} \xi, & \text{if } \upsilon_{\min_t} < \upsilon_{\min_{t+1}} \\ 0, & \upsilon_{\min_t} = \upsilon_{\min_{t+1}} \\ -\xi & \upsilon_{\min_t} > \upsilon_{\min_{t+1}} \end{cases} \tag{2}$$

Algorithm 1 describes internals of the ethics module. The inputs are U_t and U_{t+1}. To implement MA, store $\upsilon_{\min_t}$ and $\upsilon_{\min_{t+1}}$ (lines 1–2). Compare $\upsilon_{\min_t}$ and $\upsilon_{\min_{t+1}}$ to assess how action a taken in s_t affected $\upsilon_{\min_{t+1}}$ (Line 3). Generate sanction F_{t+1} (Lines 4–7). Output F_{t+1} for interaction model to combine with environmental reward r_{t+1} through reward shaping so that $r'_{t+1} = r_{t+1} + F_{t+1}$. (Line 8).

Algorithm 1: Ethics module.

Input: U_t, U_{t+1}
Output: F_{t+1}

1: $\upsilon_{\min_t} \leftarrow$ getMinExperience(U_t)
2: $\upsilon_{\min_{t+1}} \leftarrow$ getMinExperience(U_{t+1})
3: **if** $\upsilon_{\min_{t+1}} > \upsilon_{\min_t}$ **then**
4: $\quad F_{t+1} = \xi$
5: **else if** $\upsilon_{\min_{t+1}} == \upsilon_{\min_t}$ **then**
6: $\quad F_{t+1} = 0$
7: **else**
8: $\quad F_{t+1} = -\xi$
9: **end if**
10: **return** F_{t+1}

Norms Module Norms module, NM, tracks patterns of behaviour the agent learns. Norms module stores behaviours in a behaviour base and norms in a norm base. For each behaviour, it computes and stores the numerosity num, obtained from the number of times the behaviour is used, and the reward r' (described in interaction module) received from using the behaviour. The fitness of each behaviour τ is obtained from $\text{num} \cdot r'$ decayed over time. Where η is the age of the behaviour and λ is the decay rate,

$$\tau(\zeta) = \text{num} \cdot r' \cdot \lambda^{\eta} \tag{3}$$

Algorithm 2 describes the internals of the norm module. Inputs to the norm module include ν_t, a_t, r'_{t+1}, where ν_t is the precondition obtained from the agent's view of state s_t (for scalability, ν_t is a subset of s_t); a_t is the action taken in s_t. Norms module searches the behaviour base to retrieve a behaviour matching $\langle \mathsf{pre}, \mathsf{act} \rangle$ to ν_t, a_t (line 1). If there is a matching behaviour, update $\tau(\zeta)$ (lines 2–3). If there is no match, behaviour learner creates a new behaviour with ν_t, a_t and adds it to behaviour base (lines 5–6). Every $t_{\text{clip\_behaviours}}$ steps, if behaviour base exceeds the maximum capacity, behaviour base is clipped to the maximum capacity by removing the least fit behaviours (lines 8–9). Norms module compares behaviour base with norm base shared by the society and stores emerged norms in norm base (line 10).

Algorithm 2: Norms module.

Input: ν_t, a_t

1: $\zeta \leftarrow$ behaviourBase.retrieve(ν_t, a_t)
2: **if** ζ! = None **then**
3: $\quad$ behaviourBase.updateFitness(ζ)
4: **else**
5: $\quad \zeta \leftarrow$ behaviourLearner.create(ν_t, a_t)
6: $\quad$ behaviourBase.add(ζ)
7: **end if**
8: **if** t % clipNorm is 0 and len(behaviourBase) > maxLen **then**
9: $\quad$ behaviourBase.clip()
10: **end if**
11: normBase.updateEmergedNorms(behaviourBase)

Interaction Module Interaction module, DM, implements RL with deep Q network (DQN) architecture (Sutton and Barto 2018). Via DQN, RAWL·E agent learns a behaviour policy to achieve goals while promoting ethical norms. At each time step t, agent selects a batch of B random experiences to train its Q network against its target network, computing the Huber loss (Huber 1964). To prevent overfitting, every C steps weights of target network are updated to weights of the Q network θ. At each step, agent receives an observation of the environment, a vector of features $x(s)$ visible in state s, which it stores in the experience replay buffer. Each feature of $x(s)$ coresponds to a feature in the agent's DQN. With probability ϵ, agent selects an action randomly or using DQN. Using DQN, actions $a \in A$ are selected which policy $\pi(s)$ estimates will maximise expected return and help achieve goals G. Agent acts asynchronously and receives a reward from its environment r. DM obtains shaped reward F_{t+1} from EM. To encourage an agent to learn behaviours which promote ethical norms whilst pursuing goals, DM combines self-directed sanction F_{t+1} with environmental reward r_{t+1} through reward shaping so that $r'_{t+1} = r_{t+1} + F_{t+1}$. Transition $(a_t, s_t, s_{t+1}, r'_{t+1})$ is stored

in experience replay buffer. DM obtains view ν_t from state s_t and passes ν_t to NM for norm learning.

Algorithm 3 outlines the interaction module. Input environmental observation at s_t, which includes environment state, agent's resources d, and well-being $\upsilon_1, \ldots, \upsilon_n$ of all agents $ag_1, \ldots, ag_n$. Deterministic policy $\Pi(\theta, a)$ defines the agent's behaviour in s_t to output action a_t (Line 1). After acting, observe r_{t+1}, s_{t+1} (Line 2); obtain well-being vectors U_t and U_{t+1} with $\upsilon_1, \ldots, \upsilon_n$ obtained from s_t and s_{t+1} (Lines 3–4); pass U_t and U_{t+1} to EM to obtain F_{t+1} (Line 5); obtain r'_{t+1} from r_{t+1} and F_{t+1} (Line 6); update $\Pi(\theta, a)$ (Line 7); obtain ν_t from s_t (Line 8); pass ν_t to NM to learn and store behaviours and norms (Line 9).

Algorithm 3: Interaction module.

Input: s_t

1: $a_t \leftarrow \pi(s_t)$ /* *Obtain action from policy* */
2: r_{t+1}, $s_{t+1} \leftarrow$ act(a_t) /* *Perform action, observe* r_{t+1}, s_{t_+1} */
3: $U_t \leftarrow$ getWellbeing(s_t) /* *Obtain well-being* */
4: $U_{t+1} \leftarrow$ getWellbeing(s_{t+1})
5: $F_{t+1} \leftarrow$ EthicsModule(U_t, U_{t+1}) /* *Obtain sanction* */
6: $r'_{t+1} \leftarrow r_{t+1} + F_{t+1}$ /* *Shape reward* */
7: $\Pi(\theta, a) \leftarrow$ update(Π, s_t, r'_{t+1}, s_{t+1}) /* *Update policy* */
8: $\nu_t \leftarrow$ getView(s_t) /* *Obtain view of* s_t */
9: NormsModule(ν_t, a_t, r'_{t+1}) /* *Update norms module* */

4 Simulation Environment

We evaluate RAWL·E agents in a simulated harvesting scenario where they forage for berries. Cooperative behaviours may emerge, such as agents learning to throw berries to one another. To demonstrate the efficacy of modular ethical analysis, the scenario includes environmental rewards for cooperation. Figure 1 shows our harvesting environment.

4.1 Scenario

The environment represents a cooperative multi-agent scenario with a finite population of agents on a $o \times p$ grid. Time is represented in steps. At the beginning of each episode, the grid is initialised with $k = 4$ agents, and $b_{\text{initial}} = 12$ berries at random locations. An agent begins with $h_{\text{initial}} = 5.0$ health. Agents may collect berries, throw berries to other agents, or eat berries. An agent receives a gain in health $h_{\text{gain}} = 0.1$ when it eats a berry. Agent health decays $h_{\text{decay}} = -0.01$ at every time step. An agent dies if its health level reaches 0 and episode ends when all agents have died. Appendix A.2, Tables 4 and 5 provide complete list of simulation parameters and parameters for the interaction module.

Agents act asynchronously, in a different random order on each step of the simulation. At each step, each agent ag_i decides to move (north, east, south, west), eat a berry, or throw a berry to another agent ag_j if ag_i has at least $h_{\text{throw}} = 0.6$ health. When an agent has eaten a berry, a berry regrows at a random location on the grid. At each step, an agent forages for a berry in its location. An agent observes its health, its berries, distance to the nearest berry, and each agent's well-being. Well-being is represented by a function of an agent's health and number of berries it has in its bag:

$$ag_{\text{well-being}} = \frac{ag_{\text{health}} + (ag_{\text{berries}} \times h_{\text{gain}})}{h_{\text{decay}}} \tag{4}$$

For each agent, at each time step:
(1) Receive observation s_t
(2) Choose a using DQN: move (north, south, east, west), eat, throw
(3) Forage for berry; update health (h_{decay} at each step, h_{gain} if berry eaten)
(4) Receive transition: r_{t+1}, s_{t+1}, check if done
(5) Pass transition to Q network to learn
(6) Every C steps, update θ of target network
(7) Pass transition to norms module, update norm base
(8) Check health, if agent has died remove from the grid

For testing, we run each simulation $e = 2000$ times, with each simulation running until all agents have died, or a maximum of $t_{\max} = 50$ steps. We select these numbers empirically. Agents clip behaviour every $t_{\text{clip\_behaviours}}$ steps, clip norm base every $t_{\text{clip\_norms}}$ steps, and check for emerged norms every step. Table 1 lists the norm parameters.

Table 1: Norm parameters.

Parameter	Description	Value
$t_{\text{clip\_behaviours}}$	Clip behaviour base frequency	10.0
$t_{\text{clip\_norms}}$	Clip norm base frequency	5.0

4.2 Society Types for Evaluation

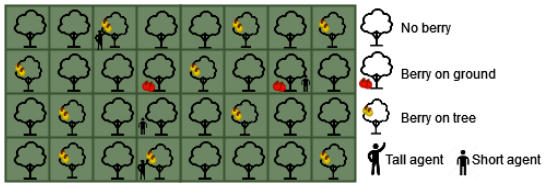
We implement two types of agent societies for evaluation.
Baseline Cooperative: DQN A society consists of standard DQN agents who do not implement an ethics module but receive environmental rewards for cooperative behaviour. DQN agent makes decisions according to its observations and expected reward.
RAWL·E: Rawlsian DQN A society of RAWL·E agents act in ways that promote Rawlsian ethics. RAWL·E agent makes decisions according to its observations and expected reward, considering the well-being of all agents.

4.3 Environmental Rewards

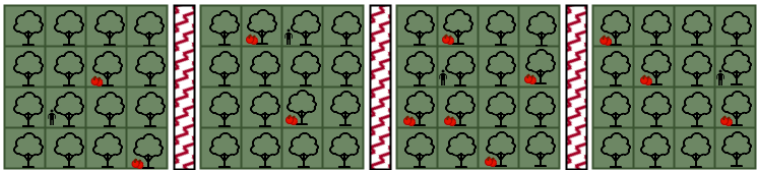
An agent receives a positive reward if it forages for a berry in a location where a berry is growing, if it eats a berry when it has berries in its bag, or if it survives to the end of the episode. An agent receives a negative reward if it attempts to eat or throw a berry to another agent when it doesn't have any, or if it dies. Agent deaths are included in raw rewards to provide incentives for societies to survive.

Self-directed sanction of a RAWL·E agent is 0.4 if the minimum experience was improved, –0.4 if the agent could have improved the minimum experience but did not (i.e., if an action was available to improve the minimum experience but the agent chose an alternative action), and 0 otherwise.

To avoid obvious results by giving RAWL·E agents additional rewards, we normalise rewards between baseline and



(a) Capabilities harvest. Agents move freely through the grid but can only harvest certain berries. Some berries are on the ground, only visible by short agents. Others are in trees, only visible by tall agents. Agents can learn to throw berries to one another across the grid.



(b) Allotment harvest. Agents are assigned a certain allotment in a community garden. Agents can only harvest berries within their allotment. Each allotment has a different amount of berries that grow there. Agents can learn to throw berries to agents in other allotments.

Figure 1: Harvesting environment. (a) Capabilities harvest scenario explores how agents learn to identify and reach desired berries while considering the well-being of the society. (b) Allotment harvest scenario explores how agents learn to harvest within their desired areas while considering the well-being in the society.

RAWL·E agents such that RAWL·E agents receive lower raw rewards. This allows for fairer comparison between societies. Table 2 summarises rewards an agent receives; in Appendix B, Table 6 displays the complete list of rewards.

Table 2: Rewards received by an agent. To avoid obvious results by giving RAWL·E agents more rewards, we normalise rewards between baseline and RAWL·E agents.

Action	Baseline	RAWL·E
Survive episode	1.0	1.0
Eat berry	1.0	0.8
Forage where berry is	1.0	0.8
Throw berry to others	0.5	0.5
Die	−1.0	−1.0
Improve minimum experience	0.0	0.4
Did not improve minimum experience	0.0	−0.4

4.4 Metrics and Hypotheses

Emerged norms $\mathcal{N}$ describe the standards of expected behaviour in a society. To evaluate $\mathcal{N}$, we examine cooperative norms which emerge by their fitness and numerosity. We assess the effects of those norms on societal outcomes with the following metrics and hypotheses.

Variables To quantitatively assess societal outcomes, for each simulation run, we record the following variables:
V_1 ($ag_{\text{well-being}}$) Number of days an agent has left to live, a function of number of berries an agent carries and their current health (Equation 4).
V_2 (ag_{resource}) Number of berries eaten by an agent.

Metrics To assess fairness on an individual and at societal level, we compute the metrics M_1 (inequality) and M_2 (minimum experience) on each variable.
M_1 (inequality) Gini index across the society. Lower is better. 0 denotes perfect equality; 1 denotes perfect inequality.
M_2 (minimum experience) Lowest individual experience across the society. Higher is better.

To assess the sustainability of the society, we compute the metrics M_3 (social welfare) and M_4 (robustness).

M_3 (social welfare) How much society as a whole gains (Mashayekhi et al. 2022). Higher is better.
M_4 (robustness) Length of episode. Higher is better.

Appendix C includes further description of the metrics.

Hypotheses We evaluate the following hypotheses. Null hypotheses for each indicate no difference.
H_1 (minimum experience) Norms emerging in RAWL·E society lead to higher minimum individual experience.
H_2 (inequality) Norms emerging in RAWL·E society lead to lower inequality.
H_3 (social welfare) Norms emerging in RAWL·E society lead to higher social welfare.
H_4 (robustness) Norms emerging in RAWL·E society lead to higher robustness.

For each hypotheses, we test the significance and compute effect size. For significance, we conduct Mann-Whitney U test which is a non-parametric test for comparing two independent groups (Mann and Whitney 1947). We use Mann-Whitney U because the sample size k is small. $p < 0.01$ indicates significance. For effect size, we compute Cohen's d which assesses the magnitude of difference between means, standardised by the pooled standard deviation (Cohen 1988), calculated as $\frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}}$, where <0.2 (negligible), [0.2,0.5) (small), [0.5,0.8) (medium), and ≥0.8 (large).

5 Experimental Results

To evaluate the behaviour of RAWL·E agents, we run agents in two experiment scenarios with different demonstrations of unequal resource allocation. For testing, we run $e = 2000$ episodes, with each episode running until $t_{\max} = 50$, or until the agents die. For qualitative analysis, we examine the emerged norms and actions promoted. For quantitative analysis, we examine fairness and sustainability metrics.

5.1 Emerged Norms

RAWL·E agent's norms model learns emerging norms from patterns of behaviour. To evaluate these norms, we run e episodes for each society and store $\mathcal{N}$ from each episode. At each step, agents compare behaviour bases and store norms repeated by 90% of agents in shared norm base $\mathcal{N}$.

We observe that in both harvest scenarios, RAWL·E agents learn more cooperative norms of throwing berries than the baseline society, such as:

```
IF <high health, medium berries, low
    neighbour well-being> THEN <throw>
```

To evaluate $\mathcal{N}$ over e episodes, we examine the numerosity num obtained from the times the norm is used, and fitness τ (Equation 3) of cooperative norms. We find that RAWL·E agents learn cooperative norms with higher fitness and use cooperative norms more, indicated by higher numerosity. Appendix D.1, Table 7 provides additional details of emerged cooperative norms. Appendix D.2 provides the complete list of emerged norms.

5.2 Simulation

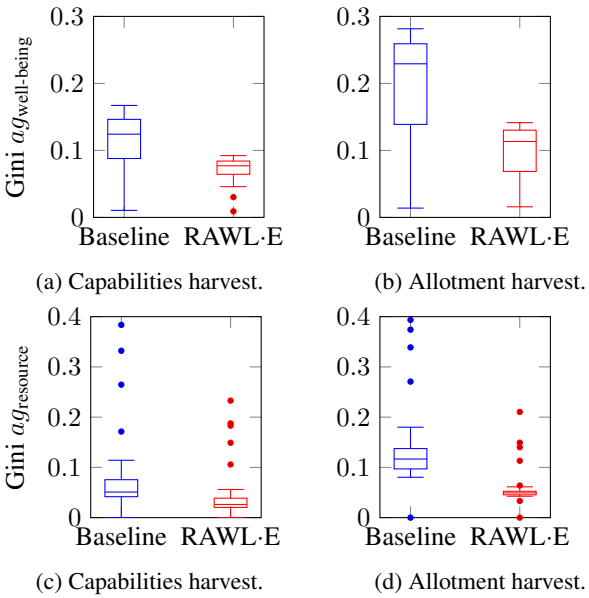
To quantitatively assess how ethical the normative system is, we analyse fairness and sustainability metrics of social welfare, inequality, minimum experience, and robustness. Table 3 summarises results for the allotment harvest; Appendix E includes additional results. We find that the results are consistent across both scenarios with method agent societies having higher social welfare, lower inequality, higher minimum experience, and higher robustness.

Table 3: Comparing ag_{resource}, inequality, minimum experience, and robustness of baseline and RAWL·E societies in allotment harvest scenario. Grey highlight indicates best results with significance at $p < 0.01$.

Metrics	Variable	Mean $\bar{x}$		Cohen's d
		Baseline	RAWL·E	
M_1 Inequality	$ag_{\text{well-being}}$	0.20	0.10	1.58
	ag_{resource}	0.14	0.06	1.32
M_2 Minimum experience	$ag_{\text{well-being}}$	7.18	10.82	3.09
	ag_{resource}	3.79	4.50	0.27
M_3 Social welfare	$ag_{\text{well-being}}$	51.50	59.80	0.64
	ag_{resource}	18.94	20.60	0.14
M_4 Robustness		47.36	48.19	0.11

H_1 (inequality) We find that RAWL·E societies have lower inequality, indicated by a lower Gini index, in both scenarios. Inequality is especially apparent in the allotment harvest for $ag_{\text{well-being}}$, where $\bar{x} = 0.2$ for the baseline society and $\bar{x} = 0.1$ for RAWL·E. We reject the null hypothesis corresponding to H_1 as $p < 0.01$ for $ag_{\text{well-being}}$ and ag_{resource}; the effect is large (1.58 for $ag_{\text{well-being}}$; 1.32 for ag_{resource}). Figure 2 compares Gini index for each society.

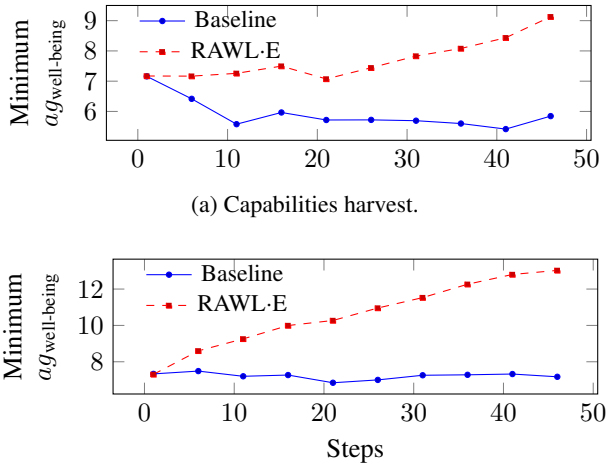
H_2 (minimum experience) RAWL·E societies have higher minimum individual experience than baseline agents in both scenarios. The largest effect (3.09) is on $ag_{\text{well-being}}$ in the allotment harvest, with $\bar{x} = 10.82$ in RAWL·E and $\bar{x} = 7.18$ for baseline. For $ag_{\text{well-being}}$, we reject the null hypothesis corresponding to H_2 as $p < 0.01$. For ag_{resource}, we cannot reject the null hypothesis in as $p > 0.01$. Figures 3 illustrate results for each society for $ag_{\text{well-being}}$.



(a) Capabilities harvest. (b) Allotment harvest.

(c) Capabilities harvest. (d) Allotment harvest.

Figure 2: Comparing Gini index of $ag_{\text{well-being}}$ and ag_{resource} for e. Lower Gini in RAWL·E indicates lower inequality.



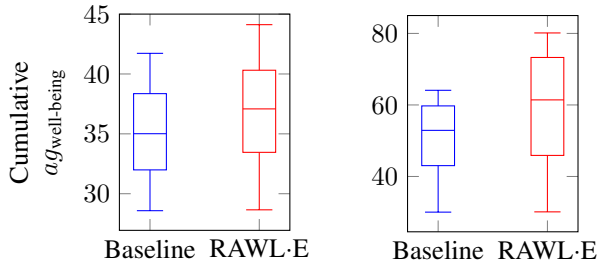
(a) Capabilities harvest.

(b) Allotment harvest.

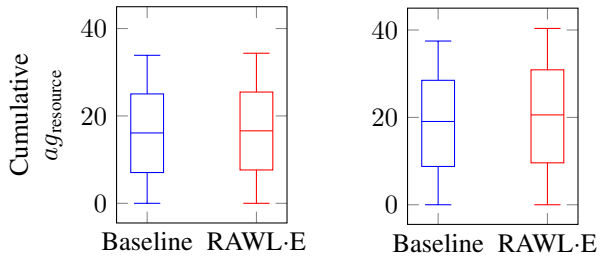
Figure 3: Minimum $ag_{\text{well-being}}$ over t_{max} steps summed for e, normalised by step frequency. RAWL·E yields higher minimum well-being.

H_3 (social welfare) RAWL·E yields higher social welfare. For $ag_{\text{well-being}}$, the allotment harvest yields $\bar{x} = 59.80$ for RAWL·E and $\bar{x} = 20.60$ for baseline which has a medium effect (0.64). We reject the null hypothesis corresponding to H_3 for $ag_{\text{well-being}}$ ($p < 0.01$), the difference, however, for ag_{resource} is not significant. Figure 4 displays these results.

H_4 (robustness) RAWL·E societies survive longer ($\bar{x} = 48.19$ in allotment) than baseline societies ($\bar{x} = 47.36$ in allotment) indicating higher robustness. We reject the null



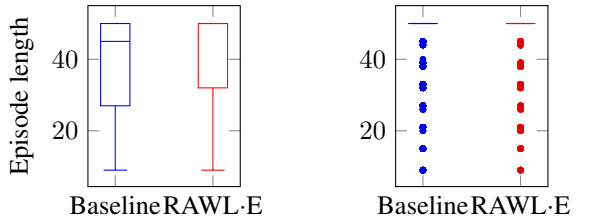
(a) Capabilities harvest. (b) Allotment harvest.



(c) Capabilities harvest. (d) Allotment harvest.

Figure 4: Cumulative $ag_{\text{well-being}}$ and ag_{resource} of each society over $t_{\max}$ steps summed for e, normalised by step frequency. Societies of RAWL·E agents have higher well-being and cumulative resource consumption.

hypothesis corresponding to H_4 as $p < 0.01$; the effect is negligible. Figures 5a and 5b show results for each society.



(a) Capabilities harvest. (b) Allotment harvest.

Figure 5: Days survived for e. Societies of RAWL·E agents survive for longer, indicating higher robustness.

Summary of Findings Our results support our hypotheses. Our main findings are: (1) in a society of RAWL·E agents, social welfare is improved, indicated by higher cumulative resource consumption, (2) inequality is reduced, indicated by a lower Gini index, (3) minimum individual experience is higher than the baseline; the combination of reduced inequality and improved minimum individual experience suggests that RAWL·E societies are fairer, and (4) RAWL·E societies survive longer, indicating higher robustness. Together, these results suggest RAWL·E agents promote the emergence of norms which improve fairness and social welfare, thereby promoting considerate behaviour, further leading to a more sustainable society.

We observe that results are better (higher fairness, social welfare, and robustness) for RAWL·E than baseline in both scenarios. However, the difference is more apparent in the allotment harvest than capabilities harvest. We attribute this difference to the fact that in the capabilities harvest agents are in a more confined space than the allotment harvest, and must navigate around one another to reach berries.

Threats to Validity Threats arise from the simplicity of our scenarios. While this abstraction limits real-world applicability, our focus is on demonstrating the operationalisation of normative ethics rather than capturing realism. To address this threat, we present our agent architecture decoupled from the environment. Also, using shaped rewards to operationalise ethics offers an adaptable method compatible with various RL algorithms and diverse scenarios.

6 Discussion and Conclusion

Developing agents that behave in ways that promote ethical norms is crucial for ethical MAS. Operationalising principles from normative ethics in individual decision making helps address the problem of deriving an ought from an is. Our results show that, compared to societies of baseline agents who don't implement normative ethics, RAWL·E agents societies have higher social welfare, and are more fair by higher minimum experience and reduced disparity.

Directions and Key Takeaways Applying normative ethics presents challenges, and there is often disagreement on the subject (Moor 2006). Conflicts may arise when different principles promote different actions (Robinson 2023). Additionally, the application of a principle may lead to unintuitive outcomes or fail to promote one action over another (Guinebert 2020). Utilising a variety of principles in reasoning is beneficial to examine scenarios from different perspectives, improving the amplitude of ethical reasoning. Directions include operationalising a variety of principles, and investigating circumstances in which principles conflict.

We utilise rewards to promote learning ethical behaviour when not all states can be known in advance. However, modifying rewards combines different objectives in a single numerical scale, allowing implicit comparisons between outcomes (Nashed, Svegliato, and Blodgett 2023). Directions include combining promotion of ethical behaviour with explicit prevention of unethical outcomes.

The scenarios we implement are abstracted to demonstrate how the method can be implemented. Operationalising normative ethics provides a mechanism to systematically assess the rightness and wrongness of actions in a range of situations (Binns 2018). Applying our method to more complex and real world scenarios, with a range of different RL algorithms, is another direction for future work.

Reproducibility Our codebase is publicly available (Woodgate, Marshall, and Ajmeri 2024). Appendices A–E provide additional details, including computing infrastructure, parameter selection, a complete list of environmental rewards, further descriptions of metrics, a complete set of emerged norms, and additional details on simulation results.

7 Acknowledgments

JW thanks EPSRC Doctoral Training Partnership Grant No. EP/W524414/1 for the support. NA thanks UKRI EPSRC Grant No. EP/Y028392/1: AI for Collective Intelligence (AI4CI) for the support.

References

Agrawal, R.; Ajmeri, N.; and Singh, M. P. 2022. Socially Intelligent Genetic Agents for the Emergence of Explicit Norms. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 10–14. Vienna: IJCAI.

Ajmeri, N.; Guo, H.; Murukannaiah, P. K.; and Singh, M. P. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 16–24. Auckland: IFAAMAS.

Anavankot, A. M.; Cranefield, S.; and Savarimuthu, B. T. R. 2023. Towards Norm Entrepreneurship in Agent Societies. In *Advances in Practical Applications of Agents, Multi-Agent Systems, and Cognitive Mimetics. The PAAMS Collection*, 188–199. Switzerland: Springer.

Anderson, M.; Anderson, S. L.; and Armen, C. 2004. Towards Machine Ethics. In *AAAI-04 Workshop on Agent Organizations: Theory and Practice*, 1–7. San Jose: AAAI.

Balakrishnan, S.; Bi, J.; and Soh, H. 2022. SCALES: From Fairness Principles to Constrained Decision-Making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 46–55. Oxford: ACM.

Bengio, Y. 2012. Practical recommendations for gradient-based training of deep architectures. *CoRR*, abs/1206.5533.

Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAccT)*, volume 81, 149–159. New York: PMLR.

Chen, Y. F.; Everett, M.; Liu, M.; and How, J. P. 2017. Socially aware motion planning with deep reinforcement learning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 1, 1343–1350. Vancouver: IEEE.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 2nd edition.

Dell'Anna, D.; Dastani, M.; and Dalpiaz, F. 2020. Runtime Revision of Sanctions in Normative Multi-Agent Systems. *Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 34(2): 1–54.

Dignum, V. 2021. The Myth of Complete AI-Fairness. In Tucker, A.; Henriques Abreu, P.; Cardoso, J.; Pereira Rodrigues, P.; and Riaño, D., eds., *Artificial Intelligence in Medicine*, 3–8. Online: Springer.

Dong, S.; Li, C.; Yang, S.; An, B.; Li, W.; and Gao, Y. 2024. Egoism, utilitarianism and egalitarianism in multi-agent reinforcement learning. *Neural Networks*, 178: 106544.

Endriss, U. 2013. Reduction of Economic Inequality in Combinatorial Domains. In *Proceedings of the 12th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 175–182. Minnesota: IFAAMAS.

Grupen, N. A.; Selman, B.; and Lee, D. D. 2022. Cooperative Multi-Agent Fairness and Equivariant Policies. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 9350–9359. Online.

Guinebert, S. 2020. How do moral theories stand to each other? *Zeitschrift für Ethik und Moralphilosophie*, 3(2): 279–299.

Guo, Y.; Wang, B.; Hughes, D.; Lewis, M.; and Sycara, K. 2020. Designing Context-Sensitive Norm Inverse Reinforcement Learning Framework for Norm-Compliant Autonomous Agents. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 618–625. Online: IEEE.

Huber, P. J. 1964. Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics*, 35: 492–518.

Jing, S.; and Doorn, N. 2020. Engineers' Moral Responsibility: A Confucian Perspective. *Science and Engineering Ethics*, 26(1): 233–253.

Kim, T. W.; Hooker, J.; and Donaldson, T. 2021. Taking Principles Seriously: A Hybrid Approach to Value Alignment in Artificial Intelligence. *JAIR*, 70: 871–890.

Kittock, J. E. 1995. Emergent Conventions and the Structure of Multi-Agent Systems. In *Lectures in complex systems: The proceedings ofthe 1993 complex systems summer school, Santa Fe Institute Studies in the Sciences of Complexity Lecture Volume VI*, 507–521. Santa Fe Institute, Addison-Wesley.

Leben, D. 2020. Normative Principles for Evaluating Fairness in Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 86–92. New York: ACM.

Levy, P.; and Griffiths, N. 2021. Convention Emergence with Congested Resources. In Rosenfeld, A.; and Talmon, N., eds., *Multi-Agent Systems*, 126–143. Cham: Springer International Publishing.

Mann, H. B.; and Whitney, D. R. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1): 50 – 60.

Maranhão, J.; Casini, G.; Pigozzi, G.; and van Der Torre, L. 2022. Normative change: an AGM approach. *Journal of Applied Logics - IfCoLoG Journal of Logics and their Applications*, 9(4): 855–920.

Mashayekhi, M.; Ajmeri, N.; List, G. F.; and Singh, M. P. 2022. Prosocial Norm Emergence in Multiagent Systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 17(1–2): 3:1–3:24.

Moor, J. 2006. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, 21: 18–21.

Morris-Martin, A.; De Vos, M.; and Padget, J. 2019. Norm Emergence in Multiagent Systems: A Viewpoint Paper. *Autonomous Agents and Multi-Agent Systems (JAAMAS)*, 33(6): 706–749.

Muñoz, A.; Billsberry, J.; and Ambrosini, V. 2022. Resilience, robustness, and antifragility: towards an appreciation of distinct organizational responses to adversity. *International Journal of Management Reviews*, 24: 181–187.

Murukannaiah, P. K.; Ajmeri, N.; Jonker, C. M.; and Singh, M. P. 2020. New Foundations of Ethical Multiagent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1706–1710. Auckland: IFAAMAS. Blue Sky Ideas Track.

Murukannaiah, P. K.; and Singh, M. P. 2020. From Machine Ethics to Internet Ethics: Broadening the Horizon. *IEEE Internet Computing*, 24(3): 51–57.

Nardin, L. G.; Balke-Visser, T.; Ajmeri, N.; Kalia, A. K.; Sichman, J. S.; and Singh, M. P. 2016. Classifying Sanctions and Designing a Conceptual Sanctioning Process Model for Socio-Technical Systems. *The Knowledge Engineering Review (KER)*, 31: 142–166.

Nashed, S.; Svegliato, J.; and Zilberstein, S. 2021. Ethically Compliant Planning within Moral Communities. In *Proceedings of the 4th AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 188–198. Virtual Event: ACM.

Nashed, S. B.; Svegliato, J.; and Blodgett, S. L. 2023. Fairness and Sequential Decision Making: Limits, Lessons, and Opportunities. *ArXiv*, abs/2301.05753: 1–15.

Neufeld, E. A.; Bartocci, E.; Ciabattoni, A.; and Governatori, G. 2022. Enforcing ethical goals over reinforcement-learning policies. *Ethics and Information Technology*, 24(4): 43.

Oldenburg, N.; and Zhi-Xuan, T. 2024. Learning and Sustaining Shared Normative Systems via Bayesian Rule Induction in Markov Games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1510–1520. Auckland: IFAAMAS.

Ong, Y. C.; Protopapas, N.; Yazdanpanah, V.; Gerding, E. H.; and Stein, S. 2024. Fair and efficient ride-scheduling: a preference-driven approach. *Journal of Simulation*, 1(1): 1–17.

Rawls, J. 1958. Justice as Fairness. *The Philosophical Review*, 67(2): 164–194.

Rawls, J.; and Kelly, E. I. 2001. *Justice As Fairness : A Restatement.* Online Access with DDA: YBP Pick and Choose. Cambridge: Belknap Press.

Robinson, P. 2023. Moral disagreement and artificial intelligence. *AI and Society*, 38(3): 1–14.

Savarimuthu, B. T. R.; Cranefield, S.; Purvis, M. A.; and Purvis, M. K. 2013. Identifying prohibition norms in agent societies. *Artificial Intelligence and Law*, 21(1): 1–46.

Shoham, Y.; and Tennenholtz, M. 1997. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94(1): 139–166. Economic Principles of Multi-Agent Systems.

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

Svegliato, J.; Nashed, S. B.; and Zilberstein, S. 2021. Ethically Compliant Sequential Decision Making. *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, 35(13): 11657–11665.

Tuomela, R. 1995. *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford: Stanford University Press.

Tzeng, S.-T.; Ajmeri, N.; and Singh, M. P. 2022. Fleur: Social Values Orientation for Robust Norm Emergence. In *Proceedings of the International Workshop on Coordination, Organizations, Institutions, Norms and Ethics for Governance of Multi-Agent Systems (COINE)*, 185–200. Virtual: Springer.

Vinitsky, E.; Köster, R.; Agapiou, J. P.; Duéñez-Guzmán, E. A.; Vezhnevets, A. S.; and Leibo, J. Z. 2023. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, 2(2): 1–14.

Woodgate, J.; and Ajmeri, N. 2022. Macro Ethics for Governing Equitable Sociotechnical Systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 1824–1828. Online: IFAAMAS. Blue Sky Ideas Track.

Woodgate, J.; and Ajmeri, N. 2024. Macro Ethics Principles for Responsible AI Systems: Taxonomy and Directions. *ACM Computing Surveys*, 56(11): 1–37.

Woodgate, J.; Marshall, P.; and Ajmeri, N. 2024. Codebase for Operationalising Rawlsian Ethics for Fairness in Norm-Learning Agents. https://doi.org/10.5281/zenodo.14520386.

Wright. 1963. *Norm and Action: A Logical Enquiry*. New York: Humanities.

Yaman, A.; Leibo, J. Z.; Iacca, G.; and Wan Lee, S. 2023. The emergence of division of labour through decentralized social sanctioning. *Proceedings of the Royal Society B: Biological Sciences*, 290(2009).

Yu, C.; Zhang, M.; and Ren, F. 2014. Collective Learning for the Emergence of Social Norms in Networked Multiagent Systems. *IEEE Transactions on Cybernetics*, 44(12): 2342–2355.

Zimmer, M.; Glanois, C.; Siddique, U.; and Weng, P. 2021. Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, 12967–12978. Online: PMLR.

A Details of Experimental Setups

This section provides details about the computing infrastructure and hyperparameter selection.

A.1 Computing Infrastructure

We conducted the simulation experiments on a workstation with Intel Xeon Processor W-2245 (8C 3.9 GHz), 256GB RAM, and Nvidia RTA A6000 48GB GPU.

A.2 Hyperparameter Selection

Table 4 lists the simulation parameters and range of values tried per parameter. Results are consistent across the range of values tried, with societies of RAWL·E agents having higher social welfare, fairness, and robustness than societies of baseline agents.

Table 4: Parameters for simulation experiments.

Description	Parameter	Range Tried	Final Value
Capabilities grid size	$o_{\text{capabilities}} \times p_{\text{capabilities}}$	$\{4 \times 4, 8 \times 4\}$	8×4
Allotment grid size	$o_{\text{allotment}} \times p_{\text{allotment}}$	$\{8 \times 4, 16 \times 4\}$	16×4
Number of agents	k	$\{2, 4\}$	4
Initial number of berries	b_{initial}	$\{8, 12, 16\}$	12
Initial health of agent	h_{initial}	$\{5.0, 10.0\}$	5.0
Health gain from eating berry	h_{gain}	$\{0.1, 1.0\}$	0.1
Health decay	h_{decay}	$\{-0.01, -0.1\}$	-0.01
Minimum health to throw	h_{throw}	$\{0.5, 0.6, 1.0\}$	0.6
Number of episodes	e	$\{1000, 2000\}$	2000
Maximum steps in episode	t_{max}	$\{20, 50\}$	50

Table 5 lists the interaction module parameters and range of values tried per parameter. We select these parameters empirically, with reference to literature (Bengio 2012).

B Environmental Rewards

To avoid obvious results by giving RAWL·E agents additional rewards, we normalise rewards between baseline and RAWL·E agents such that RAWL·E agents receive lower raw rewards. This allows for fairer comparison between societies. Table 6 displays the complete list of rewards received by baseline and RAWL·E agents.

C Metrics

Here, we provide further details about the metrics used to evaluate societies of RAWL·E and baseline agents.

To assess the fairness of a society, we compute M_1 (inequality) and M_2 (minimum experience).

M_1 (inequality) Examining the inequality of a society to assess fairness is supported by the principle of egalitarianism, which states that disparity amongst members should be minimised (Murukannaiah et al. 2020). We use the Gini index for inequality as it is well studied and has been used previously in MAS (Endriss 2013).

M_2 (minimum experience) Examining the minimum individual experience to assess fairness is justified by Rawlsian ethics, which argues that those who benefit the least should be prioritised (Rawls and Kelly 2001).

The fairest society will have the lowest inequality and highest minimum individual experience. However, the notion of fairness is abstract and achieving perfect fairness is challenging, if not impossible (Dignum 2021). We thus aim for satisfactory outcomes that promote equitable systems, which have a higher goal of fairness, but might not be perfect.

To assess the society's sustainability, we compute the metrics M_3 (social welfare) and M_4 (robustness).

M_3 (social welfare) Measuring social welfare (how much society as a whole gains (Mashayekhi et al. 2022)) is supported by the principle of utilitarianism, which states that ethical actions are those which maximise utility (Ong et al. 2024).

M_4 (robustness of society) Robustness relates to the degree a society is sensitive to exogenous influence, exhibited as the ability to resist and withstand adversity (Muñoz, Billsberry, and Ambrosini 2022).

D Results for Emerged Norms

D.1 Results for Emerged Cooperative Norms

A norm is emerged when it is adopted by over 90% of the population. In societies of RAWL·E agents, the cooperative norms which emerge have higher fitness and are used more frequently, indicated by higher numerosity. Table 7 summarises the results for emerged cooperative norms.

D.2 System of Emerged Norms

Figures 6 and 7 display the system of emerged norms $\mathcal{N}$ in baseline and RAWL·E societies. To obtain $\mathcal{N}$, we run $e = 2000$ episodes for $t_{\text{max}} = 50$ steps and track the norms which emerge in each episode. We combine $\mathcal{N}$ for e to obtain the list of all norms which emerge. Norms with "throw" consequent are cooperative, as throwing is an act of agents helping one another. To distill specific norms into generalised rules, we aggregate antecedent conditions which produce the same outcome. For example, all instances of the condition "no berries" result in the consequent "move". Therefore, specific norms are aggregated to the generalised rule of:

```
IF <no berries> THEN <move>
```

Table 5: Parameters of the Interaction Module.

Description	Parameter	Range Tried	Final Value	Criterion
Batch size	B	{32, 64, 128}	64	Training time
Iteration for updating weights of target network	C	{1000, 100, 50}	50	Test performance
Probability of exploration	ϵ	0.9–0.0	0.0	Test performance
Learning rate	α	{0.01, 0.001, 0.0001}	0.0001	Test performance
Number of hidden units	Hn	{32, 64, 128}	128	Test performance
Number of hidden layers	Hl	1–3	2	Test performance

Table 6: Rewards received by an agent. Rewards are normalised between baseline and RAWL·E agents to avoid obvious results by giving RAWL·E agents more rewards.

Action	Baseline	RAWL·E
Survive episode	1.00	1.00
Eat berry	1.00	0.80
Forage where berry is	1.00	0.80
Throw berry	0.50	0.50
Try to eat without berries	−0.20	−0.10
Try to throw without berries	−0.20	−0.10
Try to throw without sufficient health	−0.20	−0.10
Try to throw without recipient	−0.20	−0.10
Die	−1.00	−1.00
Improve minimum experience	0.00	0.40
No difference to minimum experience	0.00	0.00
Did not improve minimum experience	0.00	−0.40

In both scenarios, we observe that in societies of RAWL·E agents cooperative norms which emerge are more generalised than cooperative norms emerging in societies of baseline agents. For example, in Figure 6b a general norm emerges:

```
IF <high health> THEN <throw>
```

In contrast, the cooperative norms which emerge in baseline societies in Figure 6a are more specialised than in RAWL·E societies. This indicates that cooperative norms in RAWL·E societies cover a wider range of circumstances.

E Simulation Results

Table 8 provides additional details of the simulation results.

We observe that societies of RAWL·E agents have significantly lower inequality than societies of baseline agents for both $ag_{\text{well-being}}$ and ag_{resource}. The effect is medium to large in both scenarios: 1.58 for $ag_{\text{well-being}}$ and 0.63 for ag_{resource} in the capabilities harvest; 1.58 for $ag_{\text{well-being}}$ and 1.32 for ag_{resource} in the allotment harvest.

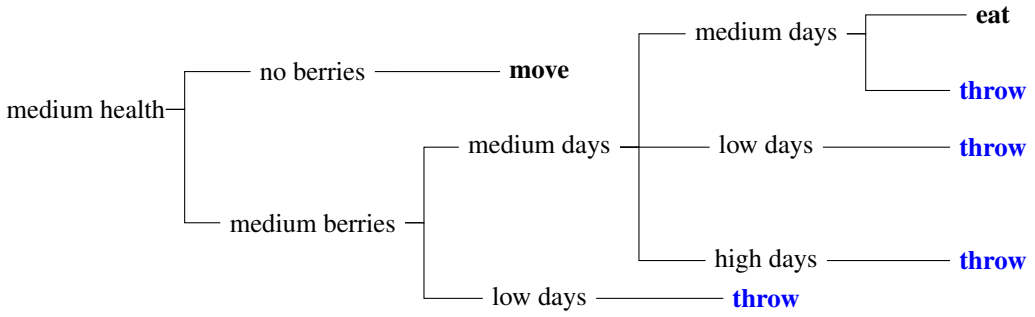
For minimum experience, societies of RAWL·E agents show significantly higher results than baseline societies for $ag_{\text{well-being}}$ in both scenarios with a large effect, however, the minimum experience is not significantly different for ag_{resource}.

Social welfare is significantly higher for $ag_{\text{well-being}}$ in societies of RAWL·E agents than baseline societies in the allotment harvest with a medium effect of 0.64. The difference in social welfare is not significant for ag_{resource}.

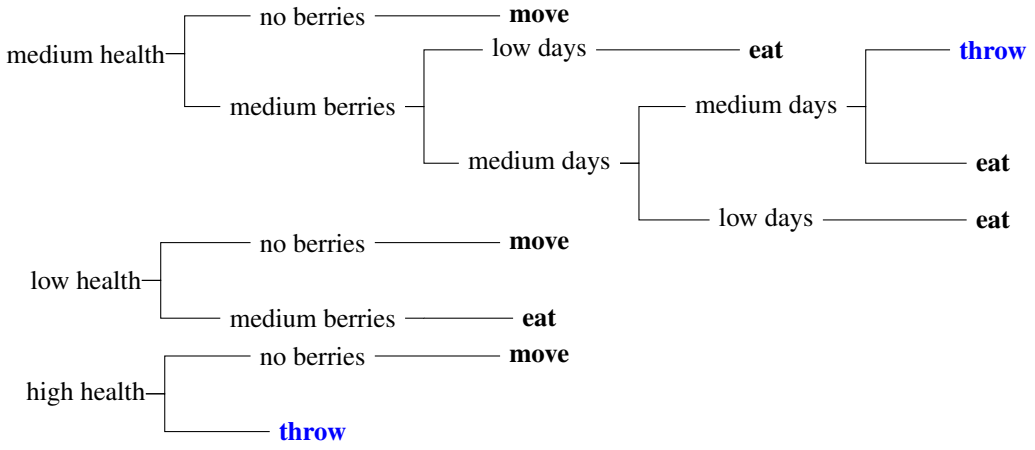
Further, in both scenarios, societies of RAWL·E agents are more robust than baseline societies; however, the effect is negligible (0.18 in capabilities harvest and 0.11 in allotment harvest).

Table 7: Comparing fitness and numerosity of cooperative norms of baseline and RAWL·E societies in capabilities harvest and allotment harvest scenarios. Grey highlight indicates best results with significance at $p < 0.01$.

Scenario	Metrics	Mean $\bar{x}$		Standard deviation σ		Cohen's d
		Baseline	RAWL·E	Baseline	RAWL·E	
Capabilities Harvest	Fitness	25.61	57.62	41.28	83.44	0.3
	Numerosity	14.61	26.58	16.4	20.54	0.25
Allotment Harvest	Fitness	46.5	64.34	84.59	92.51	0.26
	Numerosity	26.32	41.94	24.57	32.28	0.26

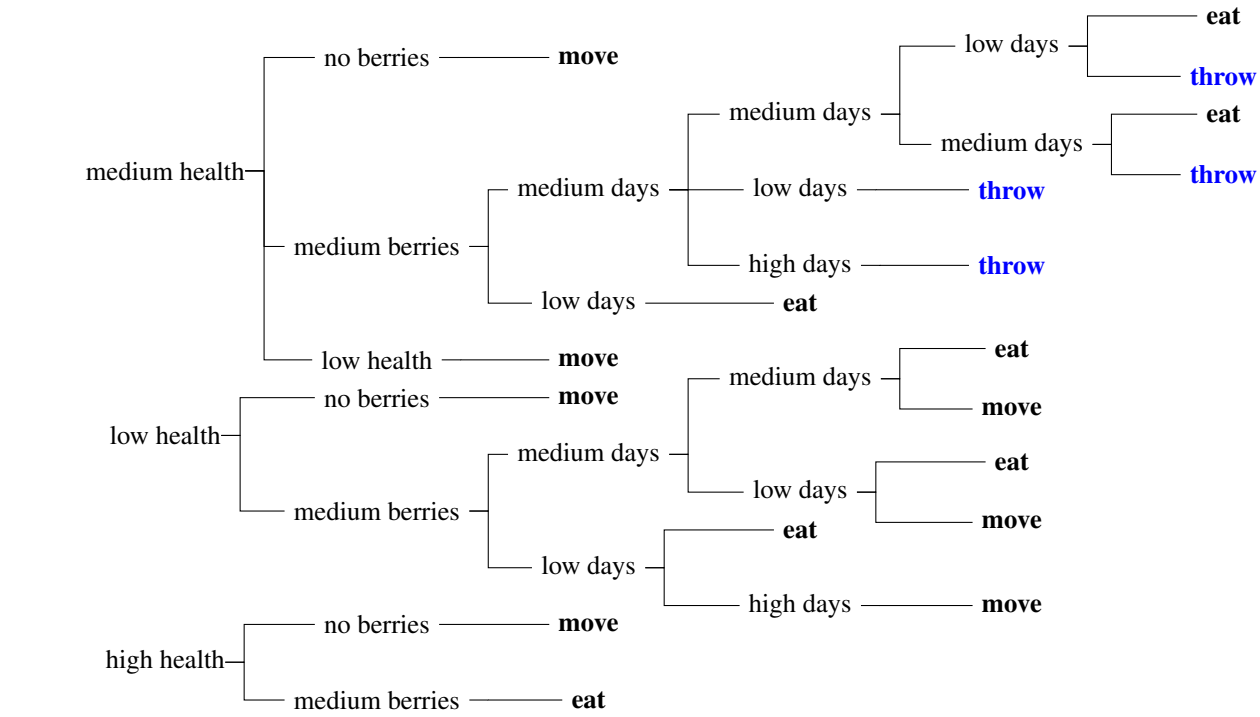


(a) Baseline norms

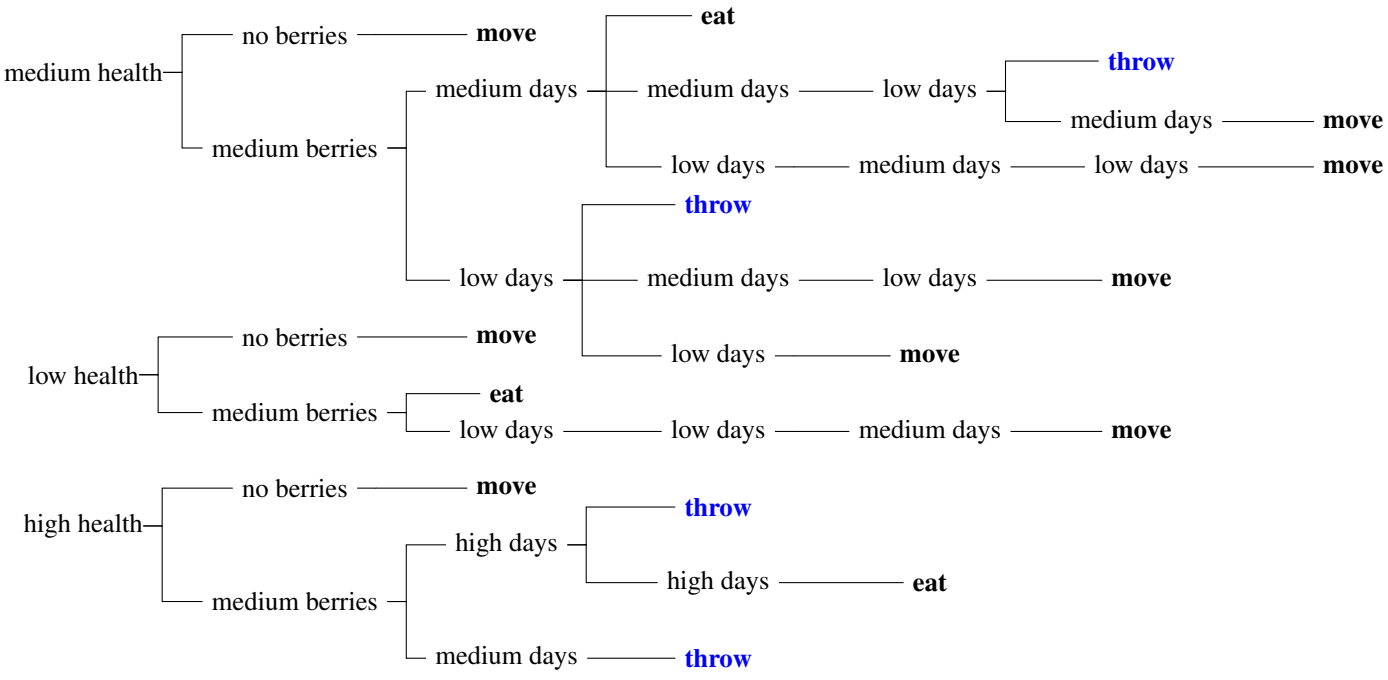


(b) RAWL·E norms

Figure 6: $\mathcal{N}$ for capabilities harvest over e_{epochs}. blue highlights cooperative norms. In societies of RAWL·E agents, more generalised cooperative norms emerge than in baseline societies. For example, in Figure 6b the general norm of "IF high health THEN throw" emerges.



(a) Baseline norms



(b) RAWL·E norms

Figure 7: $\mathcal{N}$ for allotment harvest over e_{epochs}. blue highlights cooperative norms. In societies of RAWL·E agents, cooperative norms which emerge are more generalised than in baseline societies.

Table 8: Comparing ag_{resource}, inequality, minimum experience, and robustness of baseline and RAWL·E societies in capabilities harvest and allotment harvest scenarios. Grey highlight indicates best results with significance at $p < 0.01$.

Scenario	Metrics	Variable	Mean $\bar{x}$		Standard deviation σ		Cohen's d
			Baseline	RAWL·E	Baseline	RAWL·E	
Capabilities Harvest	M_1 Inequality	$ag_{\text{well-being}}$	0.12	0.07	0.04	0.02	1.58
		ag_{resource}	0.09	0.04	0.1	0.05	0.63
	M_2 Minimum experience	$ag_{\text{well-being}}$	5.81	7.69	0.36	0.62	3.71
		ag_{resource}	3.62	3.98	2.41	2.52	0.15
	M_3 Social welfare	$ag_{\text{well-being}}$	35.25	37.05	3.62	4.39	0.45
		ag_{resource}	16.52	16.96	10.41	10.43	0.04
	M_4 Robustness		38.07	40.5	14.15	13.48	0.18
Allotment Harvest	M_1 Inequality	$ag_{\text{well-being}}$	0.2	0.1	0.08	0.04	1.58
		ag_{resource}	0.14	0.06	0.08	0.03	1.32
	M_2 Minimum experience	$ag_{\text{well-being}}$	7.18	10.82	0.23	1.65	3.09
		ag_{resource}	3.79	4.5	2.43	2.73	0.27
	M_3 Social welfare	$ag_{\text{well-being}}$	51.5	59.8	9.9	15.41	0.64
		ag_{resource}	18.94	20.6	11.35	12.28	0.14
	M_4 Robustness		47.36	48.19	7.67	6.94	0.11