Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)

Ethics, Prosperity and Society: Moral Evaluation Using Virtue Ethics And Utilitarianism

Aditya Hegde* , **Vibhav Agarwal*** and **Shrisha Rao**
International Institute of Information Technology - Bangalore, Bangalore, India
{aditya.shridhar, vibhav.agarwal}@iiitb.org, shrao@ieee.org

Abstract

Modelling ethics is critical to understanding and analysing social phenomena. However, prior literature either incorporates ethics into agent strategies or uses it for evaluation of agent behaviour. This work proposes a framework that models both, ethical decision making as well as evaluation using virtue ethics and utilitarianism. In an iteration, agents can use either the classical Continuous Prisoner's Dilemma or a new type of interaction called moral interaction, where agents donate or steal from other agents. We introduce moral interactions to model ethical decision making. We also propose a novel agent type, called virtue agent, parametrised by the agent's level of ethics. Virtue agents' decisions are based on moral evaluations of past interactions. Our simulations show that unethical agents make short term gains but are less prosperous in the long run. We find that in societies with positivity bias, unethical agents have high incentive to become ethical. The opposite is true of societies with negativity bias. We also evaluate the ethicality of existing strategies and compare them with those of virtue agents.

1 Introduction

Philosophers have enquired into the moral underpinnings of human conduct since the ancient Greek era [Rogers, 1937]. The study of ethics deals with the very definition of right and wrong, and the formulation and application of principles that define behaviour affecting other living organisms. Normative ethics is the branch of ethics that discusses when an action is right or wrong and the questions that arise when one performs an action. Normative ethical theories can be divided into three categories: Kantian, consequentialism, and virtue ethics [van Roojen, 2001]. Kantian or deontological ethics evaluate an action based on a set of moral rules (deontology) rather than the consequences of the action. On the other hand, consequentialism emphasises the effects (consequences) of actions. An action that increases or brings about something that is considered 'good' by the ethical theory, is deemed to be morally acceptable [Alexander and Moore, 2016]. Virtue ethics emphasises the inherent moral character (virtue) of actions. While all three approaches accommodate virtue, consequences and rules, they differ in what they consider to be fundamental [Hursthouse and Pettigrove, 2018].

Modelling ethics provides insights into the relationships between macro-properties of society and the corresponding ethical theory, in addition to augmenting our toolkit for modelling social phenomena [Doloswala, 2014; Korb *et al.*, 2010; Lim *et al.*, 2008]. One of the first instances of combining computer simulations with ethics was the work of Danielson [1992], who coined the term 'artificial morality' and introduced moral agents in an attempt to improve performance in games like the Iterated Prisoner's Dilemma.

Ethics can be used to perform a moral evaluation of agent behaviour and simulations [Korb *et al.*, 2010; Cointe *et al.*, 2016; Wang *et al.*, 2017]. For example, Korb et al. [2010] simulate evolving worlds where agents can pass on traits to the next generation and evaluate the effects of altruism, rape and abortion using utilitarianism, a type of consequentialist ethics. Ethics can also be used to constrain the behaviour of agents to make ethical decisions [Wiegel and van den Berg, 2009; Gaudou *et al.*, 2014]. Lasquety-Reyes [2018] details approaches to incorporate popular ethical theories like utilitarianism, feminist care ethics, Kantian ethics, etc., in agent-based models.

There are efforts focused on the study of ethics in artificial intelligence [Bostrom and Yudkowsky, 2014], on how AI systems should be designed for morality [Boyles, 2017], and on ideas for moral evaluation of machine actions [Wallach, 2010]. However, our approach for understanding the effects of ethics in society is unique and has not been directly anticipated.

This work considers both ethical decision-making as well as evaluation of the same. We also introduce a new agent type, called a *virtue agent*, which can be used to instantiate agents with different levels of ethics. Through simulations, we analyse the relationship between ethics, prosperity and society. We find that unethical agents make short-term gains but are worse off in the long run. Moreover, the prosperity of agents increases steeply with their level of ethics. Our simulations indicate that societies with positivity bias (societies that emphasise rewarding ethical actions more than penalising unethical ones) lead to convergent trends where unethical agents

*These authors contributed equally to this work.

have high incentive to become ethical and ethical agents have incentive to maintain their position. The opposite is true for societies with negativity bias, where unethical agents might become more unethical for short term gains. We also evaluate the ethicality of existing CPD strategies and compare them to those of virtue agents.

In an iteration in our model, agents can use either the classical Continuous Prisoner's Dilemma (CPD) [Killingback and Doebeli, 2002; Verhoeff, 1993] or a new type of interaction called a *moral interaction*. The latter is introduced to model ethical decision-making and provides agents with the opportunity to either steal from or donate to other agents. The prosperity of agents is captured through a resource parameter which is updated through the CPD payoffs and moral interactions. Agents also maintain an opinion value of other agents in the simulation and update it post-interaction. These opinion updates are modelled after the principles of virtue ethics and utilitarianism. Moral interactions are followed by a broadcast of the details of the interaction to a large fraction of agents in the simulation to model the spread of information about beneficent and maleficent actions in the real world. The agents receiving the broadcast update their opinions after evaluating the ethicality of the interaction.

Virtue agents are parameterised by the agent's level of ethics. They serve as a powerful and flexible tool to instantiate populations with varying levels of ethics. The framework is modular in the sense that one can include virtue agents along with other classical agent types.

2 Framework

We use a cellular automaton model where agents are represented as cells on a grid. The set of all agents $\mathcal{A}_i$ is denoted by $\mathbb{A}$. In a simulation, the agents participate in iterated interactions with their Moore neighbours of range 1 [Weisstein, 2005]. An agent $\mathcal{A}_i$'s Moore neighbours are denoted by $\mathcal{N}(\mathcal{A}_i)$. Specifically, an iteration consists of every agent $\mathcal{A}_i$ interacting with a randomly-chosen $\mathcal{A}_j \in \mathcal{N}(\mathcal{A}_i)$. We refer to a pair of interacting agents as $\mathcal{A}_0$ and $\mathcal{A}_1$, and other agents in the simulation as $\mathcal{A}_j$.

The interactions themselves are governed by a combination of agent and model parameters, some of which are static while others may change post interaction. Central to the interactions are the opinion and resource parameters. Every agent $\mathcal{A}_0$ maintains an opinion about every other agent $\mathcal{A}_1$ in the simulation, denoted by $\Psi_{\mathcal{A}_0}(\mathcal{A}_1)$ where $0 \leq \Psi_{\mathcal{A}_0}(\mathcal{A}_1) \leq 1$. We use $\mathcal{A}_0$'s opinion about $\mathcal{A}_1$ as $\mathcal{A}_0$'s perception of $\mathcal{A}_1$'s ethicality. The average opinion about $\mathcal{A}_0$ across all agents is computed as shown in (1) and is used as the reputation of $\mathcal{A}_0$.

$$\frac{\sum_{x \in \mathbb{A} \setminus \{\mathcal{A}_0\}} \Psi_x(\mathcal{A}_0)}{|\mathbb{A}| - 1} \tag{1}$$

Every agent is also associated with a resource parameter. An agent $\mathcal{A}_0$'s resource is an integer denoted by $r_{\mathcal{A}_0}$ and is initialized with the same value for all agents at the start of the simulation. The resource is then updated based on agent interactions. We interpret $r_{\mathcal{A}_0}$ as $\mathcal{A}_0$'s prosperity in the society.

Agents can use two types of interactions, either the Continuous Prisoner's Dilemma (CPD) or a moral interaction, each of which has been described in further detail in Sections 2.1 and 2.2 respectively. The behaviour of an agent in both types of interactions is determined by its strategy. In an iteration, an agent chooses a moral interaction over CPD with a probability θ. In addition to ethical decision making, our framework also provides a way to evaluate the ethicality of agent actions as discussed in Section 2.3. The model parameters are summarised in Table 1.

2.1 Continuous Prisoner's Dilemma

The well known Prisoner's Dilemma (PD) involves two agents choosing among two possible options, cooperate or defect, which in turn affects the payoffs each agent receives. This game played repeatedly over multiple iterations is known as the Iterated Prisoner's Dilemma (IPD) [Axelrod and Hamilton, 1981] and allows agents to make choices based on previous interactions. However, both games are discrete in nature, and thus are unsuitable to model more complex scenarios.

The Continuous Prisoner's Dilemma (CPD) [Killingback and Doebeli, 2002; Verhoeff, 1993] overcomes this drawback by allowing agents to choose any level of cooperation between 0 (complete defection) and 1 (complete cooperation). Consider two agents $\mathcal{A}_0$ and $\mathcal{A}_1$ with cooperation levels c_0 and c_1 respectively. Their payoffs are a linear interpolation between the discrete game payoffs as shown in (2) [Verhoeff, 1993].

$$\begin{aligned} p_{\mathcal{A}_0}(c_0, c_1) &= c_0 c_1 R + c_0 \bar{c_1} S + \bar{c_0} c_1 T + \bar{c_0} \bar{c_1} P \\ p_{\mathcal{A}_1}(c_0, c_1) &= c_1 c_0 R + c_1 \bar{c_0} S + \bar{c_1} c_0 T + \bar{c_1} \bar{c_0} P \end{aligned} \tag{2}$$

$$\text{where:} \quad \bar{x} = 1 - x$$

R, S, T, P in (2) are the discrete payoffs in the standard PD as shown below.

		Agent $\mathcal{A}_1$: 1	Agent $\mathcal{A}_1$: 0
Agent $\mathcal{A}_0$	1	(R, R)	(S, T)
	0	(T, S)	(P, P)

Iterated games require $2R > T + S$ and $T > R > P > S$. We use the **donation game** variant of the payoff matrix where $R = (\alpha - \beta), T = \alpha, S = -\beta$ and $P = 0$ for any two positive integers $\alpha > \beta$ [Hilbe *et al.*, 2013]. In our framework, an agent $\mathcal{A}_0$ starts a CPD interaction by choosing a random neighbour $\mathcal{A}_1$. $\mathcal{A}_0$ and $\mathcal{A}_1$ then choose their cooperation levels depending on their individual strategies. The agents then receive their payoffs and update their parameters, most important of which is their opinion of each other. Post-interaction opinion updates are discussed in detail in Section 2.3.

2.2 Moral Interaction

In real life, ethical choices manifest in only a small proportion of social interactions with higher stakes [Kidder, 2009]. While CPD interactions can incorporate opinion when agents choose their cooperation levels, it does not provide a straightforward approach to model ethical decision-making.

We thus introduce a new type of agent interaction called a moral interaction. During an interaction, $\mathcal{A}_0$ opts for a moral interaction over a CPD interaction with a probability θ, where $0 < \theta \ll 1$, which captures the idea of moral interactions only forming a small proportion of all interactions. $\mathcal{A}_0$ then chooses a neighbour $\mathcal{A}_1$ for either donation or theft according to $\mathcal{A}_0$'s strategy. A donation involves a transfer of δ_d units of resource from $\mathcal{A}_0$ to $\mathcal{A}_1$ and a theft involves a transfer of δ_t units of resource from $\mathcal{A}_1$ to $\mathcal{A}_0$. To model this as a high-stakes interaction, we require $\delta_d, \delta_t > T$ where T is the CPD payoff described in Section 2.1. We note that unlike CPD, $\mathcal{A}_1$ is only the target of $\mathcal{A}_0$'s action and does not interact with $\mathcal{A}_0$.

A moral interaction is followed by a broadcast where a fraction γ of the agents in $\mathbb{A}$ update their opinions of $\mathcal{A}_0$. The broadcast helps capture real-world phenomena where beneficent and maleficent actions are widely known compared to commonplace interactions. The nature of the opinion updates is discussed in detail in Section 2.3.

To summarise, an agent occasionally gets the opportunity to either donate to or steal from one of its neighbours. When this opening comes up, the choice of neighbour as well as the decision to cheat or help is left up to the agent. Intuitively, a more ethical agent would opt to help its neighbours rather than cheat them.

2.3 Ethics

While moral interactions help model ethical decision making as discussed in Section 2.2, post-interaction opinion updates allow evaluation of the ethicality of agent actions. The updates are based on two major schools of ethics: virtue ethics and utilitarian ethics [van Roojen, 2001].

Virtue ethics emphasises the inherent moral nature of actions. In the context of our framework, higher cooperation levels and acts of donation can be seen as inherently moral actions. Utilitarian ethics, on the other hand, is a type of consequentialist ethics where an act that increases the global utility is considered to be ethical [Hursthouse and Pettigrove, 2018]. Since as discussed in Section 2, the resource parameter is used as the prosperity of an agent, we define global utility as the sum of the resource across all agents, i.e., $\sum_{x \in \mathbb{A}} r_x$.

As discussed in Section 2, we interpret $\Psi_{\mathcal{A}_0}(\mathcal{A}_1)$ to be $\mathcal{A}_0$'s perception of $\mathcal{A}_1$'s ethicality. Thus, a moral evaluation of $\mathcal{A}_1$ by $\mathcal{A}_0$ is modelled as a change in $\Psi_{\mathcal{A}_0}(\mathcal{A}_1)$.

We implement these ideas by using discrete opinion updates based on a threshold. Formally, let s denotes the total payoff after a CPD interaction between two agents $\mathcal{A}_0$ and $\mathcal{A}_1$, with cooperation levels c_0 and c_1 respectively. We update the opinion as shown in (3).

(3) consists of two parts, one based on virtue ethics and the other based on utilitarian ethics. As discussed above, a higher cooperation level is more ethical in the context of virtue ethics. Thus, we increase $\Psi_{\mathcal{A}_b}(\mathcal{A}_{1-b})$ by ω_v whenever $\mathcal{A}_{1-b}$'s cooperation level is above a threshold λ_v and decrease it by the same value otherwise. Similarly, we increase $\Psi_{\mathcal{A}_b}(\mathcal{A}_{1-b})$ by ω_u if the total payoff s (change in global utility) is above a threshold λ_u and decrease by the same value otherwise.

Parameter	Description	Value
α	Donation game payoff parameter	5
β	Donation game payoff parameter	2
θ	Probability of an agent receiving the opportunity to perform a moral interaction.	0.05
γ	Fraction of population that receives broadcast	0.8
δ_d	Resource donated in case of a donation action	20
δ_t	Resource stolen in case of a theft action	20
ω_d	Change in opinion upon donation action	0.01
ω_t	Change in opinion upon theft action	0.03
ω_v	Change in opinion post CPD interaction based on morality of action	0.004
ω_u	Change in opinion post CPD interaction based on utility of the interaction	0.002
λ_v	Threshold for morality based opinion update	0.5
λ_u	Threshold for utility based opinion update	0.0

Table 1: Summary of model parameters

$$\Psi_{\mathcal{A}_b}(\mathcal{A}_{1-b}) := \Psi_{\mathcal{A}_b}(\mathcal{A}_{1-b}) + \begin{cases} \omega_v + \omega_u & \text{if } c_{1-b} > \lambda_v \wedge s > \lambda_u \\ -\omega_v + \omega_u & \text{if } c_{1-b} \leq \lambda_v \wedge s > \lambda_u \\ \omega_v - \omega_u & \text{if } c_{1-b} > \lambda_v \wedge s \leq \lambda_u \\ -\omega_v - \omega_u & \text{otherwise} \end{cases} \quad (3)$$

$$\text{where:} \quad b \in \{0, 1\}$$

We adopt a similar approach to update opinions after moral interactions. However, the update is based only on virtue ethics since a moral interaction does not change global utility. If agent $\mathcal{A}_0$ opts for a moral interaction, an agent $\mathcal{A}_j$ updates its opinion of $\mathcal{A}_0$ as described in (4) where $\mathcal{A}_j$ is either the target of $\mathcal{A}_0$'s action (i.e. $\mathcal{A}_1$) or belongs to the fraction γ of the population which receives the broadcast.

$$\Psi_{\mathcal{A}_j}(\mathcal{A}_0) := \Psi_{\mathcal{A}_j}(\mathcal{A}_0) + \begin{cases} \omega_d & \text{if donation} \\ -\omega_t & \text{if theft} \end{cases} \quad (4)$$

We require $\omega_t > \omega_d$, i.e., the change in $\Psi_{\mathcal{A}_j}(\mathcal{A}_0)$ to be larger with theft than with donation, to model the notion that "bad is stronger than good" [Baumeister *et al.*, 2001]. (We also call this condition as *negativity bias*, with the opposite being *positivity bias*.) To model morality interactions as having higher stakes, we require $\omega_t, \omega_d \gg \omega_v, \omega_u$.

In (3) and (4), $\Psi_{\mathcal{A}_b}(\mathcal{A}_{1-b})$ and $\Psi_{\mathcal{A}_j}(\mathcal{A}_0)$ are clipped if an update causes either of them to exceed the valid range of $[0, 1]$.

3 Virtue Agents

In this section we introduce a new agent type called a virtue agent, which can be used to instantiate agents with different levels of ethics. Our framework requires all agent types to have the opinion and resource parameters. In addition to these, virtue agents also have a parameter ϵ for the agent's level of ethics, where $0 \leq \epsilon \leq 1$. A distinguishing aspect of the virtue agent type is that it uses the opinion of other agents

in addition to its own level of ethics when performing an action, i.e., it also incorporates the perceived ethicality of other agents. This is similar to the approach of Spencer [Smith, 1982] which holds that that social perception is tied to morality. Agent parameters are summarised in Table 2.

3.1 CPD Interaction

As discussed in Section 2.1, when an agent $\mathcal{A}_0$ opts for a CPD interaction, it chooses a random neighbour $\mathcal{A}_1$ with which it interacts. Thus, a strategy for CPD interaction for $\mathcal{A}_b$ involves outputting the cooperation level given $\mathcal{A}_{1-b}$, where $b \in \{0, 1\}$.

A virtue agent $\mathcal{A}_b$ aggregates the opinion of the other agent $\mathcal{A}_{1-b}$ from its neighbours to output the cooperation level c_b as shown in (5).

$$w_{\mathcal{A}_b}(x) = \mathcal{H}_{1,1}(\Psi_{\mathcal{A}_b}(x))$$
$$c_b = \frac{\sum_{x \in \mathcal{N}(\mathcal{A}_b) \setminus \{\mathcal{A}_{1-b}\}} w_{\mathcal{A}_b}(x)\ \Psi_x(\mathcal{A}_{1-b})}{\sum_{x \in \mathcal{N}(\mathcal{A}_b) \setminus \{\mathcal{A}_{1-b}\}} w_{\mathcal{A}_b}(x)} \quad (5)$$

where: $\mathcal{H}_{1,1} \rightarrow$ half-normal distribution obtained from the $\mathcal{N}_{1,1}$ normal distribution

It is essentially a weighted mean of the neighbour's opinion about $\mathcal{A}_{1-b}$. $\mathcal{A}_b$ weights agent x's opinion in proportion to $\mathcal{A}_b$'s perceived ethicality of x. We believe such an aggregation models our society, where our opinion is shaped by those around us, especially those who we hold in high regard [Moussaïd *et al.*, 2013; Campbell-Meiklejohn *et al.*, 2010].

Once cooperation levels are output by the individual agent strategies, the payoffs are computed using the CPD matrix and the resource parameter is updated for the interacting agents followed by the opinion updates as discussed in Section 2.

3.2 Moral Interaction

All agents get the opportunity to participate in a moral interaction instead of a CPD interaction with probability θ as described in Section 2.2. The agent's strategy for moral interactions has two outputs: the action (either donation or theft), and the neighbour on which the action is performed.

Intuitively, given the option between donation and theft, an ethical agent opts for the former. The virtue agent type models this idea by using the ϵ parameter as the probability of the agent to opt for donation over theft. Looking ahead, using ϵ as the probability gives a straightforward approach to quantify ethics as a continuous value that is in turn useful for analysing the results of simulations.

Once a virtue agent $\mathcal{A}_0$ has decided the action, the target $\mathcal{A}_1$ is chosen according to (6). When $\mathcal{A}_0$ wishes to donate, it chooses a relatively poor (low resource) neighbour of whom it has a high opinion. Similarly, $\mathcal{A}_0$ steals from a relatively rich (higher resource) neighbour of whom it has a low opinion.

$$v_{\mathcal{A}_0}(x) = \frac{\Psi_{\mathcal{A}_0}(x) + 1}{r_x + 1}$$
$$\mathcal{A}_1 = \begin{cases} \arg\max_{x \in \mathcal{N}(\mathcal{A}_0)} v_{\mathcal{A}_0}(x) & \text{if donation} \\ \arg\min_{x \in \mathcal{N}(\mathcal{A}_0)} v_{\mathcal{A}_0}(x) & \text{if theft} \end{cases} \quad (6)$$

Parameter	Description	Starting value
id	Unique identifier for every agent	
$\mathcal{N}(A)$	Set of Moore neighbours of agent $\mathcal{A}$ in the cellular automaton	
$r_{\mathcal{A}}$	Resource of agent $\mathcal{A}$	100
$\Psi_{\mathcal{A}}(\mathcal{B})$	$\mathcal{A}$'s opinion (between $[0, 1]$) about $\mathcal{B}$	0.5
ϵ	Probability of virtue agent to opt for donation over theft	

Table 2: Summary of agent parameters

It is known that higher social class is often correlated with greater unethical behaviour [Piff *et al.*, 2012], but we do not consider this aspect. Our presumption of "stealing from the rich" is based on research that indicates that criminals often focus on targets that they consider more lucrative [Vandeviver and Bernasco, 2019]. Likewise, the premise of "giving to the poor" follows from the idea that people like to make gifts which they believe will make a tangible difference, to targets they like [Cryder and Loewenstein, 2011].

Once the action as well as the target agent $\mathcal{A}_1$ is output by $\mathcal{A}_0$'s strategy, their resource is updated followed by a broadcast where a fraction γ of $\mathbb{A}$ update their opinions of $\mathcal{A}_0$ as discussed in Section 2.3.

4 Experiments and Results

Agents are placed on a toroidal grid on arbitrary cells such that no two agents are on the same cell. We configure the parameters with the default values mentioned in Tables 1 and 2 unless explicitly specified otherwise. Each simulation consists of 50 agents for a given value of the ϵ parameter. We discuss our findings from simulations in the following subsections. While we present plots for a single set of simulations, the results have been verified across different parameters and random seeds to ensure robustness.

The contribution of morality interactions towards ethical decision making is highlighted by comparison with simulations where $\theta = 0$, i.e., agents perform only CPD interactions. We note that in such simulations, the level of ethics parameter ϵ, is not used. We look at the range of agent resources at the end of 1500 iterations. When agents perform only CPD interactions, the difference in the maximum and minimum resource value is around 500 compared to a difference of around 4000 when $\theta = 0.05$. This clearly shows that the post CPD interaction opinion updates are not as large as those of moral interaction. This reinforces our discussion in Section 2.2 that moral interactions model high-stakes interactions while forming a small fraction of all interactions.

4.1 Comparing Agent Resource Across Time

We analyse the effect of the level of ethics of an agent on its resource across time by running simulations with an agent pool comprising 250 virtue agents equally distributed among 5 levels of ethics, $\epsilon \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. We find that unethical agents have higher resources initially but have significantly lower resources in the long run as seen in Figures 1a

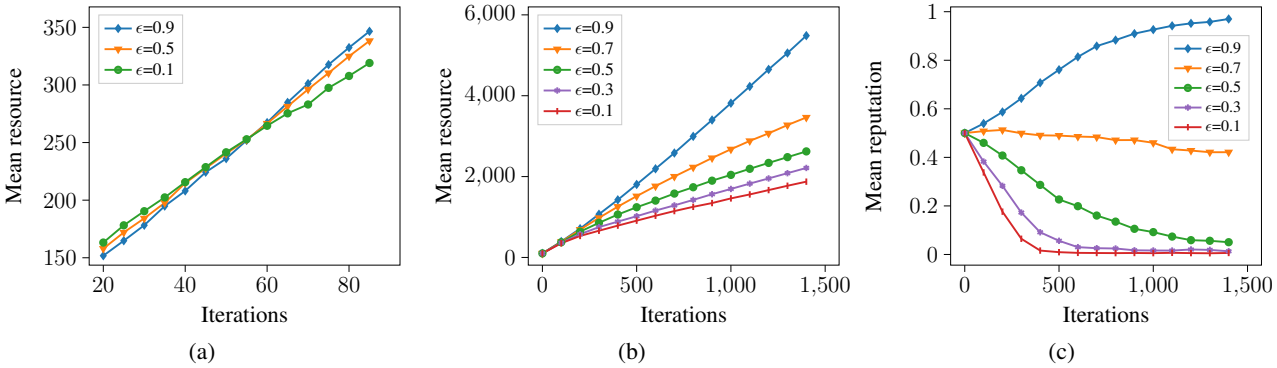


Figure 1: Mean resource and reputation across time ($\omega_d = 0.02$ & $\omega_t = 0.05$)

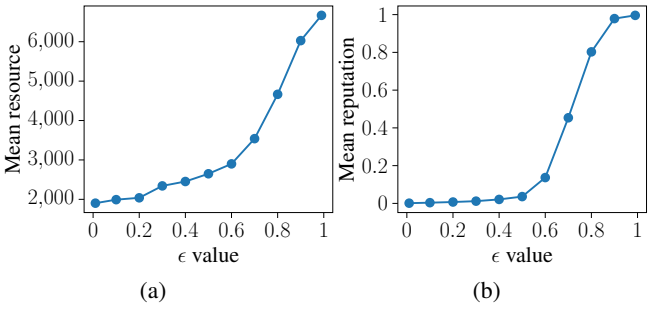


Figure 2: Mean resource and reputation at the end of 1500 iterations

ω_d	ω_t	Crossing Iteration
0.006	0.008	165
0.01	0.03	88
0.02	0.05	54

Table 3: Iteration after which ethical agents are prosperous

and 1b. For clarity, the curves corresponding to only three levels of ethics are shown in Figure 1a.

The gap between the reputations of ethical and unethical agents increases with $\omega_d + \omega_t$ since a theft action leads to a loss of ω_t in the reputation whereas a donation action leads to gain of ω_d. Thus, $\omega_d + \omega_t$ is closely related to the extent to which society favours ethical actions over unethical ones.

We observe that the time up to which unethical agents remain relatively prosperous decreases with an increase in $\omega_d + \omega_t$, as shown by Table 3, due to rapid divergence of reputation between ethical and unethical agents which in turn affects subsequent interactions. While the average resource increases every iteration due to CPD interactions, we notice that there is little change in average reputation for each group as seen in Figure 1c, indicating that the simulation has achieved stability. Thus, the difference between the prosperity of ethical and unethical agents is likely to only increase with more iterations.

These findings are in agreement with the conventional wisdom that unethical actions provide short-term payoffs but leave one worse off in the long run.

4.2 Effect of Ethics on Resources in the Long Run

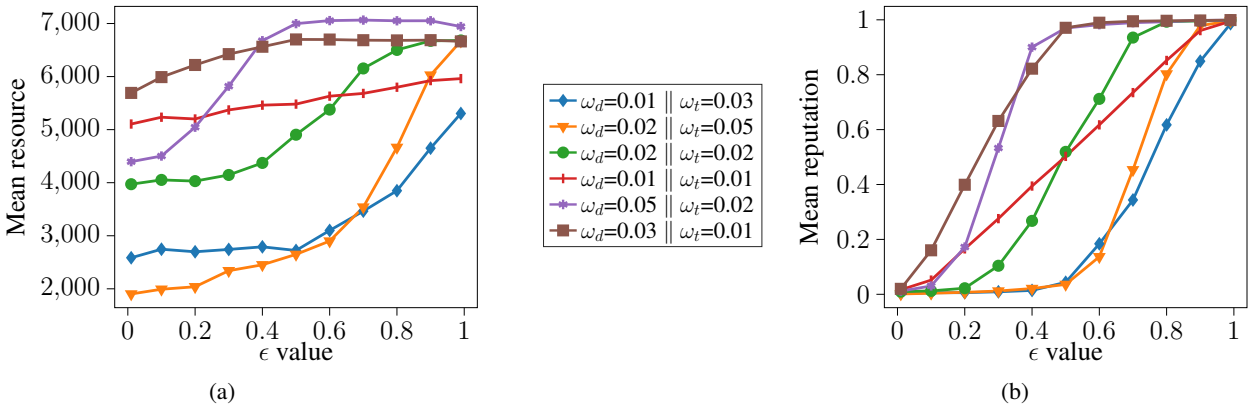
As discussed in Section 4.1, the rate of change of reputation is small after a few hundred iterations. Thus, we can reason about long-term trends by observing the state of the simulation after the reputation has stabilised. In a simulation comprising 550 virtue agents equally distributed among 11 levels of ethics, $\epsilon \in \{0.01, 0.1, 0.2, \ldots, 0.9, 0.99\}$, we observe that at the end of 1500 iterations the resource of an agent steeply increases with its level of ethics as shown in Figure 2a. Figure 2b shows a similar trend where there is a steep increase in reputation with the level of ethics.

The average reputation curve provides a possible explanation for the observed relationship between the level of ethics and average resource, since agents with higher reputations also receive higher cooperation during CPD interactions and are less likely to be chosen as targets for theft actions.

4.3 Bias in Society

As touched upon in Sections 2.3 and 4.1, ω_d and ω_t are key in determining society's perceptions of ethical and unethical actions. If $\omega_d < \omega_t$, an agent suffers a greater change (decrease) in reputation due to an unethical action than that it would have gained by performing an ethical action. We call such a configuration a *negativity bias* since unethical actions are penalised more than ethical actions are rewarded. Similarly, the configuration $\omega_d > \omega_t$ is called a *positivity bias* and $\omega_d = \omega_t$ represents a lack of bias. Thus, in Figures 3a and 3b, curves corresponding to ($\omega_d = 0.01$, $\omega_t = 0.03$) and ($\omega_d = 0.02$, $\omega_t = 0.05$) represent negativity bias; ($\omega_d = 0.03$, $\omega_t = 0.01$) and ($\omega_d = 0.05$, $\omega_t = 0.02$) represent positivity bias; and ($\omega_d = 0.01$, $\omega_t = 0.01$) and ($\omega_d = 0.02$, $\omega_t = 0.02$) reflect no bias.

To analyse the differences between a society with negativity bias, positivity bias and no bias, we run multiple simulations with different values for ω_d and ω_t using 550 virtue agents equally distributed among 11 levels of ethics, $\epsilon \in \{0.01, 0.1, 0.2, \ldots, 0.9, 0.99\}$. Figures 3a and 3b show the average resource and reputation plotted against the level of ethics after 1500 iterations for each simulation. As expected, agents with higher levels of ethics have higher reputations and resource in the long run as discussed in Section 4.2.



(a) (b)

Figure 3: Mean resource and reputation in societies with different bias

Moreover, among simulations having the same bias, the differences in resources and reputations increase with $\omega_d + \omega_t$ as discussed in Section 4.1.

We now turn our attention towards the nature of the curves, especially the rate of change of resource and reputation. From Figures 3a and 3b, we observe that when there is no bias, the rate of change is almost symmetric with respect to the level of ethics; low at the extremes and high around the middle. Negativity bias on the other hand leads to a high rate of change in the high-ethics range and a significantly lower rate of change in the low-ethics range. The opposite is true in case of positivity bias where there are minor differences in resources and reputations of high-ethics agents and significant differences in case of low-ethics agents.

We find that a society with negativity bias might lead to divergent trends where unethical agents become more unethical for short-term gains while ethical agents have the incentive to be more ethical to increase their reputation as well as resources. This is supported by our previous observation of small rates of change in reputation for unethical agents becoming more unethical but a rapid increase in reputation when ethical agents become more ethical.

Similarly, a society with positivity bias might lead to convergent trends. Here, unethical agents have the incentive of a rapidly increasing reputation as well as resource to become more ethical while there is negligible change in the resource and reputation of ethical agents (e.g., agents having a level of ethics greater than 0.4 in Figure 3b) which might lead to such agents becoming more unethical as long as it does not harm their reputation. A similar analysis shows that societies without bias do not provide any strong incentive for people to change their level of ethics.

4.4 Global Utility and Agent Composition

While Sections 4.1 to 4.3 discuss agent resource and reputation, we now turn our attention to global utility, defined as the sum of resources of all agents as discussed in Section 2.3. Global utility quantifies how prosperous the society is as a whole. We look at the relationships between global utility and the number and level of ethical agents. There are two sets of six simulations apiece, each with 100 virtue agents. One set of simulations has 90 virtue agents with 0.2 level of ethics, and the other has 90 virtue agents with 0.1 level of ethics. The remaining 10 virtue agents have an identical level of ethics chosen from $\epsilon \in \{0.4, 0.5, \ldots, 0.9\}$ across simulations. Figure 4 shows a plot of the global utility against the level of ethics of the 10 ethical agents at the end of 1500 iterations.

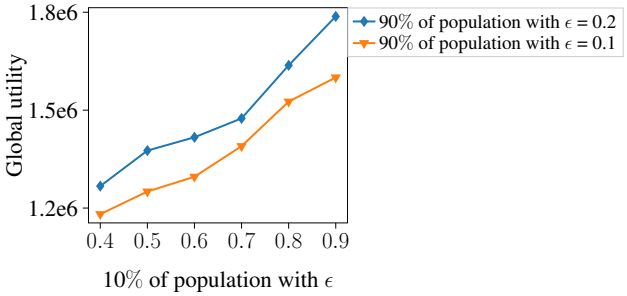


Figure 4: Global utility and level of ethics of agents in a skewed population

We find that the global utility steeply rises with the level of ethics of the ethical agents. Thus, in a society with a large proportion of unethical agents, even a small population of ethical agents leads to a significant increase in the global utility. The higher the ethics of the ethical agents, the greater is the global utility.

4.5 Ethics of Different Agent Strategies

The simulations discussed previously comprised of only virtue agents. Since the framework discussed in Section 2 is independent of the agent type, we turn to evaluating the ethicality of existing strategies like Tit-For-Tat [Axelrod and Hamilton, 1981], Suspicious-Tit-For-Tat [Boyd and Lorberbaum, 1987], Grim Trigger [Friedman, 1971] and random agents (choose a random cooperation level). These strategies have to be extended to output a target and action for morality interactions to be compatible with our framework. We do so by outputting a random neighbour as target and use the original IPD strategy for donation and theft where a donation action is interpreted as cooperate, and theft action is interpreted as defect with respect to the standard Prisoner's Dilemma.

We run a simulation comprising 250 virtue agents equally

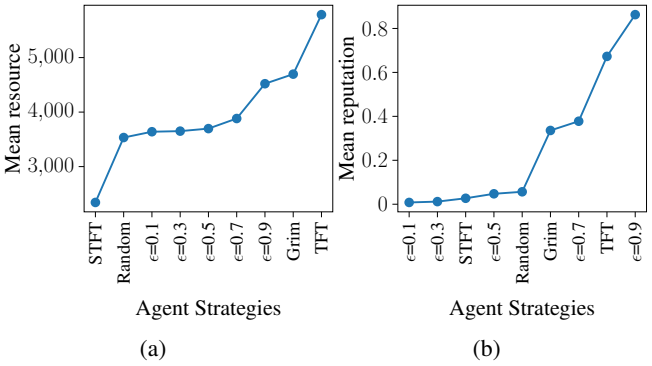


Figure 5: Comparison of existing CPD strategies

distributed among 5 levels of ethics $\epsilon \in \{0.1, 0.3, \ldots, 0.9\}$ and 50 agents each for the four alternative agent types mentioned above. The results are summarised in Figures 5a and 5b.

Intuitively, a TFT agent is expected to be more ethical than a STFT agent since the former continues to cooperate (virtuous) with the other agent until the other agent defects. We see that the simulations reinforce our intuition since TFT agents have a high reputation on average while STFT agents have a low reputation on average.

5 Conclusion

Ethics in society and in AI systems are of great contemporary interest, but the use of AI-based techniques to understand ethics in society has not hitherto been given its due. The framework we use allows for ethics to be added to agent-based models of social phenomena. The novel virtue agent type has specifically been used here to analyse the relationship between ethics, prosperity and society. However, the virtue agent can also serve as a foundation to design ethical agent behaviours in other settings and problems.

Previous studies from psychology [Rand and Nowak, 2011] as well as game theory [Dreber *et al.*, 2008; Wu *et al.*, 2009] have shown that rewards are more effective than punishments at securing cooperation. Our society however emphasises penalising unethical behaviour through law-enforcement and other punitive actions, rather than rewarding ethical behaviour [Galak and Chow, 2019]. This is in part on account of the prevalent doctrine of *retributive justice* [Walen, 2016] which sees wrongdoing as needing explicit correction through punishment, but says little about rewarding good deeds. Our results suggest that the latter may provide a stronger incentive for people to be ethical. We also see that larger rewards for ethical actions significantly diminish the transient advantages of unethical actions, and may thus help improve ethics in society.

A natural extension to the virtue agent type are agents that update their levels of ethics based on past interactions. Strategies based on optimising resources, like reinforcement learning, are particularly interesting directions for future work. Analyses of such simulations can provide deeper insights.

References

[Alexander and Moore, 2016] Larry Alexander and Michael Moore. Deontological ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016. https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/, last accessed on 01/20/2020.

[Axelrod and Hamilton, 1981] Robert Axelrod and William Donald Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, March 1981.

[Baumeister *et al.*, 2001] Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. Bad is stronger than good. *Review of General Psychology*, 5(4):323–370, 2001.

[Bostrom and Yudkowsky, 2014] Nick Bostrom and Eliezer Yudkowsky. The ethics of artificial intelligence. In Keith Frankish and William M. Ramsey, editors, *The Cambridge Handbook of Artificial Intelligence*, page 316–334. Cambridge University Press, 2014.

[Boyd and Lorberbaum, 1987] Robert Boyd and Jeffrey P. Lorberbaum. No pure strategy is evolutionarily stable in the repeated Prisoner's Dilemma game. *Nature*, 327(6117):58–59, May 1987.

[Boyles, 2017] Robert James M. Boyles. Philosophical signposts for artificial moral agent frameworks. *Suri*, 6(2), 2017.

[Campbell-Meiklejohn *et al.*, 2010] Daniel K. Campbell-Meiklejohn, Dominik R. Bach, Andreas Roepstorff, Raymond J. Dolan, and Chris D. Frith. How the opinion of others affects our valuation of objects. *Current Biology*, 20(13):1165–1170, 2010.

[Cointe *et al.*, 2016] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Ethical judgment of agents' behaviors in multi-agent systems. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, page 1106–1114, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems.

[Cryder and Loewenstein, 2011] Cynthia Cryder and George Loewenstein. The critical link between tangibility and generosity. In *Society for Judgment and Decision Making series. The science of giving: Experimental approaches to the study of charity*, pages 237–251. Psychology Press, 2011.

[Danielson, 1992] Peter Danielson. *Artificial Morality: Virtuous Robots for Virtual Games*. Routledge, 1992.

[Doloswala, 2014] Kalika Navin Doloswala. Eroding trust – An agent based model to explore how trust flows. *Australasian Marketing Journal (AMJ)*, 22(1):51–53, February 2014.

[Dreber *et al.*, 2008] Anna Dreber, David G. Rand, Drew Fudenberg, and Martin A. Nowak. Winners don't punish. *Nature*, 452(7185):348–351, 2008.

[Friedman, 1971] James W. Friedman. A Non-cooperative Equilibrium for Supergames. *The Review of Economic Studies*, 38(1):1–12, January 1971.

[Galak and Chow, 2019] Jeff Galak and Rosalind M. Chow. Compensate a little, but punish a lot: Asymmetric routes to restoring justice. *PLOS ONE*, 14(1), 2019. https: doi.org/10.1371/journal.pone.0210676, last accessed on 04/27/2020.

[Gaudou *et al.*, 2014] Benoit Gaudou, Emiliano Lorini, and Eunate Mayor. Moral Guilt: An Agent-Based Model Analysis. In Bogumił Kamiński and Grzegorz Koloch, editors, *Advances in Social Simulation*, Advances in Intelligent Systems and Computing, pages 95–106, Berlin, Heidelberg, 2014. Springer.

[Hilbe *et al.*, 2013] Christian Hilbe, Martin Andreas Nowak, and Karl Sigmund. Evolution of extortion in Iterated Prisoner's Dilemma games. *Proceedings of the National Academy of Sciences*, 110(17):6913–6918, April 2013.

[Hursthouse and Pettigrove, 2018] Rosalind Hursthouse and Glen Pettigrove. Virtue ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2018 edition, 2018. https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/, last accessed on 01/20/2020.

[Kidder, 2009] Rushworth Moulton Kidder. *How Good People Make Tough Choices Rev Ed: Resolving the Dilemmas of Ethical Living*. HarperCollins, November 2009.

[Killingback and Doebeli, 2002] Timothy Killingback and Michael Doebeli. The continuous prisoner's dilemma and the evolution of cooperation through reciprocal altruism with variable investment. *The American Naturalist*, 160(4):421–438, October 2002.

[Korb *et al.*, 2010] Kevin B. Korb, Ann E. Nicholson, and Owen Woodberry. *Evolving Ethics: The New Science of Good and Evil*. Imprint Academic, 2010.

[Lasquety-Reyes, 2018] Jeremiah Lasquety-Reyes. Computer Simulations of Ethics: the Applicability of Agent-Based Modeling for Ethical Theories. *European Journal of Formal Sciences and Engineering*, 1(2):18, July 2018.

[Lim *et al.*, 2008] Hock Chuan Lim, Rob Stocker, and Henry Larkin. Ethical Trust and Social Moral Norms Simulation: A Bio-inspired Agent-Based Modelling Approach. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 2, pages 245–251, December 2008.

[Moussaïd *et al.*, 2013] Mehdi Moussaïd, Juliane E. Kämmer, Pantelis Pipergias Analytis, and Hansjörg Neth. Social influence and the collective dynamics of opinion formation. *PLOS ONE*, 8(11):1–8, 11 2013. https://doi.org/10.1371/journal.pone.0078433, last accessed on 01/20/2020.

[Piff *et al.*, 2012] Paul K. Piff, Daniel M. Stancato, Stéphane Côté, Rodolfo Mendoza-Denton, and Dacher Keltner. Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences*, 109(11):4086–4091, 2012.

[Rand and Nowak, 2011] David G. Rand and Martin Andreas Nowak. The evolution of antisocial punishment in optional public goods games. *Nature Communications*, 2(1):1–7, 2011.

[Rogers, 1937] Reginald Arthur Percy Rogers. *A Short History of Ethics: Greek and Modern*. Macmillan and Company, Limited, 1937.

[Smith, 1982] Christopher Upham Murray Smith. Evolution and the problem of mind: Part i. Herbert Spencer. *Journal of the History of Biology*, 15(1):55–88, 1982.

[van Roojen, 2001] Mark van Roojen. Review of Three Methods of Ethics. *Philosophy and Phenomenological Research*, 62(3):721–723, 2001.

[Vandeviver and Bernasco, 2019] Christophe Vandeviver and Wim Bernasco. "location, location, location": Effects of neighborhood and house attributes on burglars' target selection. *Journal of Quantitative Criminology*, pages 1–43, 2019.

[Verhoeff, 1993] Tom Verhoeff. *A continuous version of the prisoner's dilemma*. Computing science notes. Technische Universiteit Eindhoven, 1993. https://research.tue.nl/en/publications/a-continuous-version-of-the-prisoners-dilemma, last accessed on 01/20/2020.

[Walen, 2016] Alec Walen. Retributive justice. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016. https://plato.stanford.edu/archives/win2016/entries/justice-retributive/, last accessed on 04/27/2020.

[Wallach, 2010] Wendell Wallach. Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology*, 12(3):243–250, September 2010.

[Wang *et al.*, 2017] Yijia Wang, Yan Wan, and Zhijian Wang. Using experimental game theory to transit human values to ethical AI. *ArXiv*, abs/1711.05905, 2017.

[Weisstein, 2005] Eric Wolfgang Weisstein. Moore Neighborhood, 2005. http://mathworld.worlfram.com/MooreNeighborhood.html, last accessed on 01/20/2020.

[Wiegel and van den Berg, 2009] Vincent Wiegel and Jan van den Berg. Combining Moral Theory, Modal Logic and Mas to Create Well-Behaving Artificial Agents. *International Journal of Social Robotics*, 1(3):233–242, August 2009.

[Wu *et al.*, 2009] Jia-Jia Wu, Bo-Yu Zhang, Zhen-Xing Zhou, Qiao-Qiao He, Xiu-Deng Zheng, Ross Cressman, and Yi Tao. Costly punishment does not always increase cooperation. *Proceedings of the National Academy of Sciences*, 106(41):17448–17451, 2009.