Comparative Modeling of Ethical Constructs in Autonomous Decision Making

Janvi Chhabra
International Institute of Information Technology, Bangalore
Bangalore, India
janvi.chhabra@iiitb.ac.in
Karthik Sama
International Institute of Information Technology, Bangalore
Bangalore, India
sai.karthik@iiitb.ac.in
Jayati Deshmukh
International Institute of Information Technology, Bangalore
Bangalore, India
jayati.deshmukh@iiitb.org
Srinath Srinivasa
International Institute of Information Technology, Bangalore
Bangalore, India
sri@iiitb.ac.in

***Abstract*—Computational models for ethical autonomy, are crucial for building trustworthy autonomous systems. While different paradigms of ethical autonomy are pursued, comparing and contrasting these paradigms remains a challenge. In this work, we present SPECTRA (Strategic Protocol Evaluation and Configuration Testbed for Responsible Autonomy) a general purpose multi-agent, message passing framework on top of which, different models of computational ethics can be implemented. The paper also presents our implementation of four paradigms of ethics on this framework– *deontology*, *utilitarianism*, *virtue ethics* and a recently proposed paradigm called computational transcendence. We observe that although agents have the same goal, differences in their underlying paradigm of ethics has a significant impact on the outcomes for individual agents as well as on the system as a whole. We also simulate a mixed population of agents following different paradigms of ethics and study the properties of the emergent system.**

INTRODUCTION

Autonomous agents (AAs) are getting prevalent in current times across a variety of application areas like autonomous vehicles, autonomous industrial bots, autonomous weapon systems etc. [1], [2]. The underlying algorithms of these AAs are designed such that they have a high level of autonomy and can operate with minimal or no external feedback. These AAs mostly operate in systems consisting of other AAs as well as humans and therefore their decisions and actions directly affect others. Thus, it is important that AAs take ethical aspects into consideration before taking any action. Ethical AAs have higher acceptance in societies since they can be trusted to act aligned to specific paradigms of ethics [3].

Ethics has been studied, discussed and argued for many centuries and many philosophers have proposed different paradigms of ethics. Each of these paradigms proposes its own foundational arguments in the way ethical dilemmas are approached. In realistic situations, designing ethical AAs may require some of these paradigms of ethics, or a combination of these to act as an underlying foundation for the application.

Broadly, we can classify paradigms of ethics into three overarching classes: *utilitarianism*, *deontological ethics* and *virtue ethics* [4]. *Utilitarianism* is based on resolving ethical dilemmas based on expected consequences of one's action, and aims to maximize collective utility as the underlying principle. *Deontological ethics* considers an action ethical if it follows certain rules or norms, applicable in that context. And lastly in case of *virtue ethics*, an action is deemed ethical if by doing that action, some moral virtues are manifested irrespective of actions or consequences.

Another recently proposed model of ethical agency called *Computational Transcendence* [5] argues ethical choices are a natural consequence of agents *identifying* with other agents or some larger notion. This is modeled using an *elastic* sense of self in AAs such that the utility perceived by an agent is a function of not only their own payoffs, but also payoffs accrued by all stakeholders that they identify with. With an elastic sense of self, responsible behavior is shown to be a *natural consequence* of self-interest dynamics, rather than something that conflicts with self-interest.

AAs can operate on either of these paradigms of ethics and as discussed above, each paradigm leads to an action due to an underlying argument. Although trolley problem is a hypothetical scenario, what happens in a more realistic setting? How do AAs decide which paradigm to operate on? What happens in a scenario where there are multiple AAs operating on different paradigms of ethics? What happens when some selfish adversarial agents are present in the system

along with other ethical agents? In order to answer these kind of questions, we need a testbed which can simulate and test different paradigms of ethics and in-turn compare and contrast them.

Key novel aspects presented in this paper are as follows: a) We present *SPECTRA* (Strategic Protocol Evaluation and Configuration Testbed for Responsible Autonomy) which is a computational testbed to simulate different paradigms of ethics using a message passing framework. b) Demonstrate utilitarianism, deontological ethics, virtue ethics and computational transcendence using the testbed. c) Simulated and analyzed the impact of homogeneous population of agents all following a given model of ethics. d) Simulated and analyzed the impact of heterogeneous population of agents following different paradigms of ethics. While we present simulation results for these four paradigms of ethics, the testbed is generic and can be used to simulate other paradigms of ethics as well.

COMPUTATIONAL MODELS OF ETHICS

Computational modeling of ethics can be broadly classified into three paradigmatic approaches: *consequentialism*, *deontological ethics* and *virtue ethics* [4]. We also discuss about *Computational Transcendence* [5] which leads to emergent responsible behaviour in AAs. These paradigms of ethics are summarized in Table I

Utilitarianism

Utilitarianism or consequential ethics [4], [6] is based on reasoning about the consequences of one's actions. It considers an action ethical if it leads to or maximizes overall well-being. Consequential reasoning could be based on either immediate or short-term considerations, which is called *action* consequentialism; or on long-term consequences, called *rule* consequentialism. Different models of consequentialism have been used to design a variety of computational models for responsible autonomy [7], [8], [9], [10], [11], [12], [13].

Some of the challenges of consequentialism include: difficulty in evaluating consequences, especially in open-world conditions with uncertainty. Defining utility– especially longer-term notions of "greater good" is yet another challenge. In extensive games, computing longer-term utility also comes with a high cost, and it might not even be possible to compute it in time before the agent must choose an action, which might result in agents approximating utility.

Deontology

Deontological ethics [4] considers an action ethical if the rules or principles governing that action can be considered to be universally applicable. The foundational principle for deontology was provided by Immanuel Kant called the "Categorical Imperative", which states that "act only according to that maxim whereby you can, at the same time, will that it should become a universal law" [14]. For example, one should not lie because if everyone starts lying, no one will trust anyone and human communication will lose its value.

There are two kinds of deontological models namely– agent based deontological ethics and patient based deontological ethics. Agent based theories are based on duties which are agent relative and form the core guiding rules. On the contrary, patient based theories are based on rights of agents which are agent neutral and form a qualitatively different set of guiding rules.

Some of the challenges of deontological ethics are: granularity of rules– rules must be exactly followed by agents and all the exceptions must be handled. Thus all exceptions which can occur in the system must be known in advance so that it can be handled appropriately. Conflicting rules are yet another challenge with deontological paradigms– especially in large state spaces with multiple considerations.

Virtue Ethics

Virtue ethics [4] is yet another ethical paradigm, where an action is deemed ethical if by doing that action some underlying principle or moral virtue is manifested. Virtue ethics does not look at either actions or consequences, but on agents itself and if they are displaying virtuous behaviour.

Aristotle elaborated on some of the core ideas of virtue ethics. His believed that to "sustain", people should attain some virtues in precise amount and proportion. He defined virtues as the things that "cause [their] possessors to be in a good state and to perform their functions well." (cite) In simple words, according to the role an agent has to perform, agent should try to attain the virtues which can help them to perform the role in a better way. For example, a doctor should demonstrate virtues like care, empathy etc, a soldier should demonstrate virtues like loyalty, patriotism etc in order to play their respective roles effectively. Also, the process of attaining a virtues is a continuous on-going process, thus agents should practice and demonstrate virtuous behaviour over time.

Being a virtuous agent, is not just about attainment of virtues, but also about attaining "right amount" of virtues. Extremes of any virtue signify imbalance, and thus agents should strive to maintain virtues around the *golden mean*. For example, hard work is a virtue for a researcher, however too much or too little hard work adversely affects the researcher. However, quantifying a virtue is difficult and context specific. Hence, defining a balance is difficult, it has to be an iterative process where after practicing, an agent figures what works best in that context.

Computational Transcendence

Computational transcendence [5] models an elastic sense of self in AAs such that they can identify with other agents, groups of agents and concepts in a system. It has been demonstrated that transcended agents which account for others, choose responsible actions in the presence of selfish choices. Specifically, transcended agents have two main metrics based on which they can adjust identification with others– transcendence level γ denotes the extent to which these agents care about others and semantic distance d denotes the relative importance of each aspect they identify with. Instead of having rational associations which only last as long as the association serves self-interest, transcended agents have associations of

	Deontology	Virtue Ethics	Transcendence	Utilitarianism
Description	An action is right if it is based on a moral rule or principle.	An action is right if it is based on good virtues in the right amount usually as demonstrated by virtuous agents.	An action tends to be right if one's sense of self includes other stakeholders.	An action is right if it has the best consequences for everyone.
Central Paradigm	Correct rules matter, results are irrelevant.	Focus on the attributes of the agent.	Elastic sense of self that includes the interests of other stakeholders as one's own.	Outputs matter not actions or intentions.
Criteria	Universality	Golden mean	Elastic identity	Utility maximization

TABLE I: Comparison between different models of ethics

identity, such that they associate with others because they identify with them.

Computational transcendence presents a paradigmatic departure, in this setting. Here, agents are not explicitly seeking to act virtuously. Virtuous or cooperative behavior emerges as a result of transcendence of their sense of self. This transcendence too is regulated by rational rather than normative, considerations– an agent is more likely to transcend and identify with generous environments, rather than in thrifty or unfriendly environments. Acting with transcendence does not involve any other computation for an agent, other than computing its own utility. This obviates the need for an additional layer of ethics to be added separately on an underlying agent model. For example, utilitarianism requires utility of all agents in the system to be computed, and available for each agent as common knowledge. In contrast, with computational transcendence, responsible behaviour can emerge with local knowledge. The presented model still requires some common knowledge– in the form of knowledge of payoffs of other agents with which it is interacting, in order to compute one's transcended utility.

Responsibility Dilemma

In this section, we describe an example of responsibility dilemma and discuss how different ethical paradigms would approach it. In challenging times like the COVID-19 pandemic, healthcare workers had two choices– to stay at home and protect themselves and their families or to go to hospitals and clinics to look after infected people while putting themselves at risk. Healthcare workers following different models of ethics would approach this dilemma as follows:

For a *deontic healthcare worker*, the choice should be such that it can be a universal law which can be followed by all. Thus the choice of choosing to stay at home can be catastrophic since if all healthcare workers stay at home, it would lead to dire circumstances where there would be no one to treat the infected people. Thus, a deontic healthcare worker would choose to go to the hospital.

A *utilitarian healthcare worker* tries to maximize the utility of all agents affected by the choice. Thus, going to hospital would save more number of lives than staying at home. Hence, a utilitarian agent would choose to go to the hospital.

A *virtuous healthcare worker* demonstrates virtues which are relevant in its context. In case of a healthcare worker, virtues like compassion, care, empathy and benevolence are relevant and should be upheld. These virtues can be demonstrated by a healthcare worker by being at the hospital and trying to save as many lives as possible. Thus a virtuous healthcare worker would choose to go to the hospital.

Finally, a *transcended healthcare worker* having an elastic sense of self would identify with the collective well being. Hence, this would motivate a transcended healthcare worker to choose to go to the hospital.

Thus, we note that although all types of ethical healthcare workers would choose to go to the hospital to save their patients' lives putting their families and themselves at risk, the reason why they choose to do that is different across various models of ethics.

SPECTRA FRAMEWORK

We present *SPECTRA* (Strategic Protocol Evaluation and Configuration Testbed for Responsible Autonomy) which is a multi-agent test-bed to evaluate different ethical paradigms on a common message passing framework. The primary goal is to model the dilemma of responsibility i.e. selecting among two types of choices– first which is individually beneficial for the agent but collectively adverse vs second which is sub-optimal for the individual but is good for the collective. At each step, every agent needs to choose among these two alternatives of irresponsible vs responsible choice.

We model the interactions among agents using an undirected graph, where a node represents an autonomous agent and presence of an edge between two agents represents that those agents can interact. In an interaction, an agent can assume following three roles:

(a) Sender: Initiates the interaction and sends a message to one of its direct neighbours which is intended to be sent to a receiver.
(b) Intermediate: Receives message from sender and decides whether to forward the message to receiver or drop the message.
(c) Receiver: Receives the message sent by sender via intermediate agent.

The node utility, nu and node cost, nc matrices for the sender, s and intermediate, i agents is shown in Table II. Every time, a message is sent or forwarded, it incurs a message cost, mc to the sender or intermediate agent. However, the sender only gets a message utility, mu when its message has been forwarded by the intermediate agent. Thus, there is no rational incentive for the intermediate agents to forward messages.

	Utility (nu)	Agent i	Agent s
Agent i	Forward (f)	0	mu
	Drop (d)	0	$-mu$

	Cost (nc)	Agent i	Agent s
Agent i	Forward (f)	mc	0
	Drop (d)	0	0

TABLE II: Utilities and Costs for intermediary agent (i) and source agent (s) as a result of decisions taken by intermediary

However, if intermediate agents don't forward messages, then no messages would reach intended receivers resulting in a network which does not serve any purpose.

The responsible choice is primarily to be made by the intermediate agent. First choice *drop*, d is individually beneficial as intermediate does not incur any cost on dropping a message, but it is expensive for the sender as the intended message is dropped and needs to be re-sent. On the other hand second choice *forward*, f is individually expensive as intermediate expends its resources in forwarding the message but it is good for the sender as the intended message reaches the receiver.

Thus, we need computational models of responsible agents which make choices taking collective welfare into consideration instead of just their self-interest. The test-bed models this responsibility dilemma where agents follow different paradigms of ethics and try to resolve the dilemma of whether to forward or drop a message in the network. We simulate AAs following different paradigms of ethics in a network and study the impact on the system as a whole.

Models of ethical agents

Using SPECTRA framework, we model ethical agents driven by different ethical schools of thought like deontology, virtue ethics and utilitarianism. We also model transcended agent which operates based on a recently proposed model called Computational Transcendence [5]. Agents decide to forward or drop a message based on the ethical school of thought they follow. Qualitatively, every agent differs in the following aspects:

1) Forward logic: For an intermediate agent the logic that directs whether that agent should forward or drop a particular message.
2) Stability logic: Some models of ethics have a few learnable parameters which the agents learn over multiple epochs. The system ends in a stable state once these learnable parameters of all agents in the network stabilize.

Overall block diagram of an ethical agent is shown in Figure 1. We now elaborate how theoretical constructs for responsible behaviour are modelled in agents following different models of ethics with respect to the aspects described above. The class diagram of ethical agents in this framework is presented in Figure 2. The utility obtained by sender upon its message being delivered to recipient is mu (Message Utility). While the cost an agent incurs for transmitting a message is mc (Message Cost). Both mu and mc are network parameters. For all ethical agents, $forward\_message()$ and $is\_stable()$ are the common methods and their implementation depends on the ethical paradigm they follow. While the common parameters, nu (Node Utility) and nc (Node Cost) are the total utility and total cost accrued by an agent respectively.

Deontic Agent A deontic agent operates on Immanuel Kant's "Categorical Imperative" which has been discussed in the previous section. It acts in such a way that if its actions were to become a universal law, the network would still be stable. In this framework, we don't have any fixed deontic rules to begin with. Thus, a deontic agent learns from its neighbourhood and aligns its action with them.

To model this in agents, we introduced two parameters, namely ϵ (Experience) and lr (Learning Rate). To start with, all agents forward with an ifp (Initial Forward Probability), set as a hyperparameter. The agents continue to forward with probability ifp for ϵ epochs. During these epochs they estimate how their neighbourhood forwards messages. After ϵ epochs, agents forward with a fp (Forward Probability) which they learnt. The network settles when no agent changes its forward probability. In turn, every agent has a learned experience of how the network operates and hence behaving in such a manner can be seen as each agent following the universal maxim in the setting of a message passing network. Let rr (Reach Ratio) be the ratio of messages of the agent that have reached their destination over all the messages sent by the agent. Then the learning of fp can be described as follows,

$$\begin{aligned} rr &= \frac{\text{messages reached}}{\text{messages sent}} \\ fp &= (1 - lr) * fp + (lr * rr) \end{aligned} \tag{1}$$

Virtue Agent The virtue ethics model neither focuses on the action nor its consequences but on the demonstrations of virtuous behavior by agent. In the message passing scenario, we have modeled virtue agents to exhibit the virtue of "reliability" i.e. how reliable can intermediate agents be to the sender of the message.

To model this virtue in agents, we introduced two parameters, vp (Virtue Point) and bs (Bin Size). When an agent forwards a message, it gets vp, which motivates a virtue agent to be reliable. As discussed in previous section, virtue ethics also directs an agent to strike a balance between the extremes of a virtue captured by the idea of golden mean. In the message passing network, always forwarding and always dropping every message would constitute these extreme behaviours.

The idea of golden mean has been modelled as follows: suppose each agent has a bin to accumulate vp. Agent gets vp based on their decision to forward or drop the message. This bin has a capacity, till which it can accumulate vp. The capacity of the bin is hyperparameter which will be referred as bs. At any instant, the aggregate of vp in this bin is considered vs (Virtue Score) of the agent. To decide whether to forward or drop a message, a virtue agent first computes the overall cost c incurred for forwarding a message. Accounting this cost,

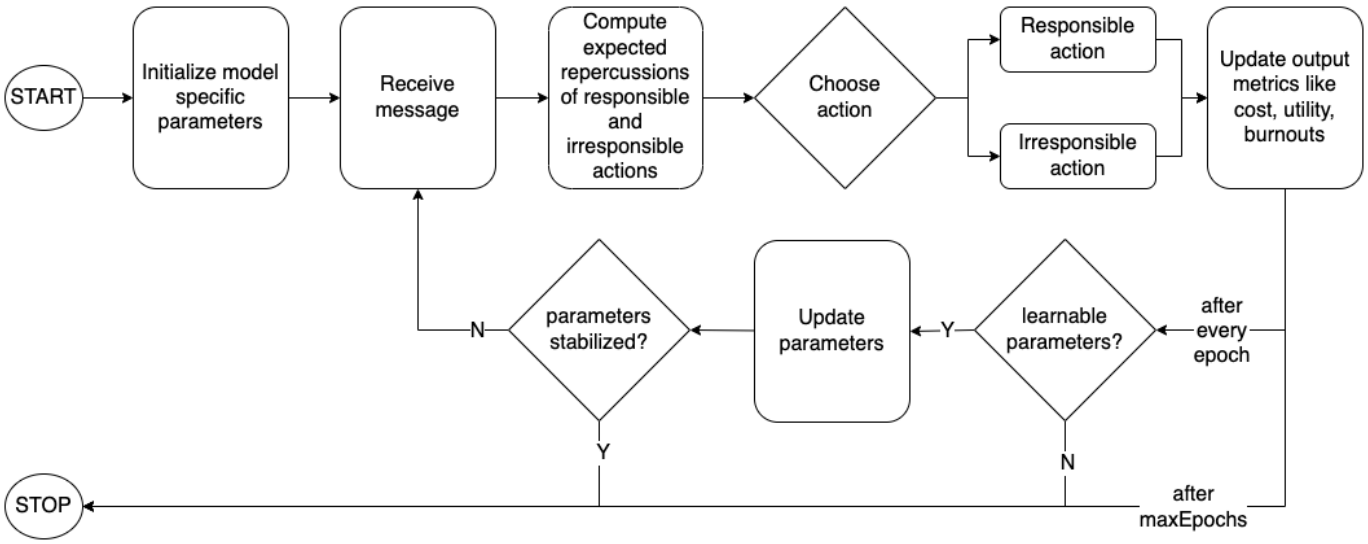


Fig. 1: Block diagram of an Ethical Agent

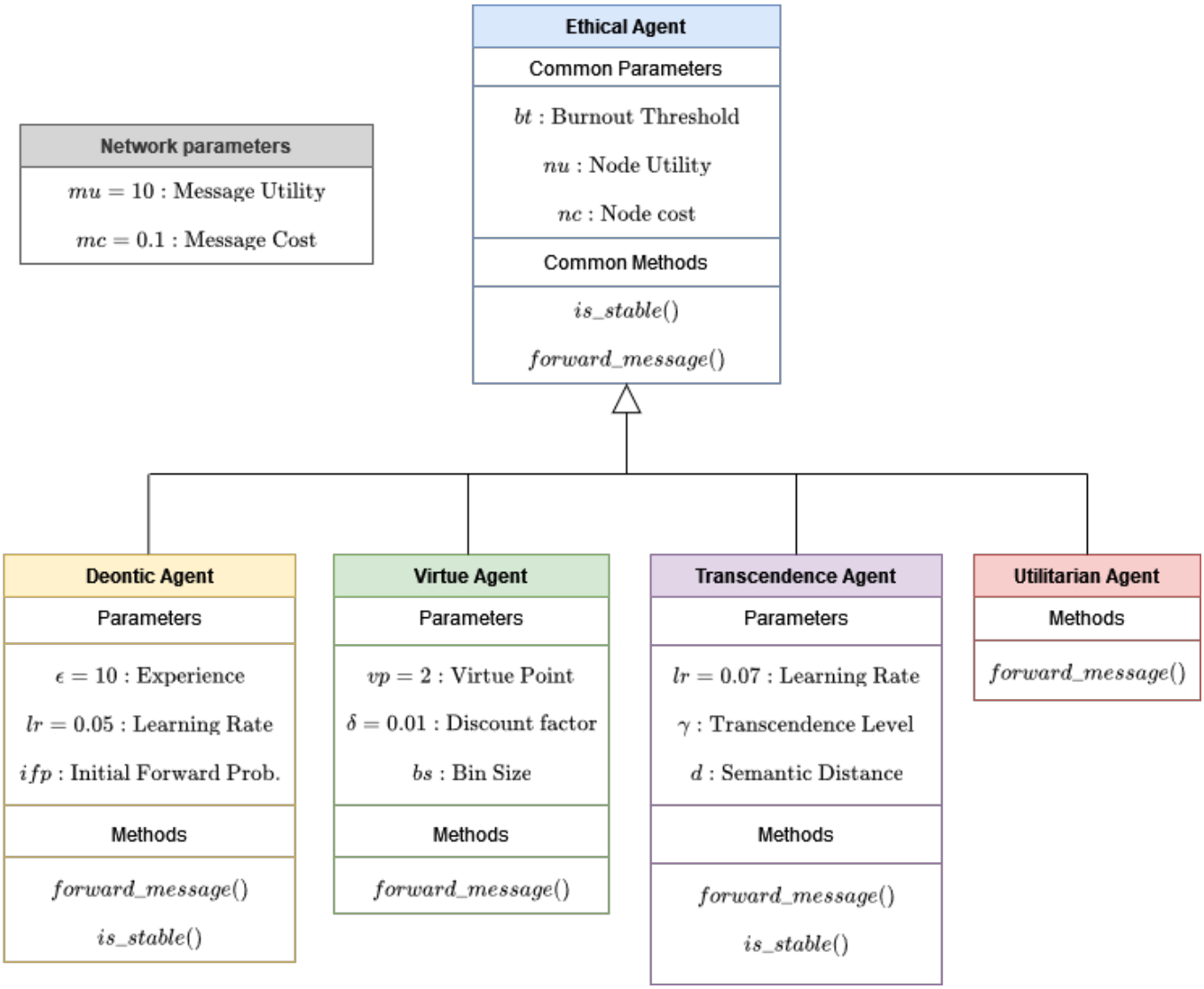


Fig. 2: Class diagram for the SPECTRA Framework

helps a virtue agent to maintain balance between the cost it incurs and virtue points it accumulates. This cost c consists of a relative historic cost (NodeCost nc - NodeUtility nu), scaled down by δ, and the current message forwarding cost mc. We express c as follows:

$$c = \delta(nc - nu) + mc \tag{2}$$

Virtue Score vs of an agent can't exceed the bin size bs. On forwarding a message, agent accumulates vp upto the limit of bs. The utility which virtue agent gets on forwarding is the net difference of the virtue points it accumulates and the cost it incurs as shown in Equation 2. On dropping the message, vp deducts from the virtue score, and agent gets utility of $-vp$. A virtue agent computes the utility it gets by forwarding or dropping a message as follows:

$$\begin{aligned} u(f) &= min(vp, bs - vs) - c \\ u(d) &= -vp \end{aligned} \tag{3}$$

Finally, the virtue agent calculates the probability to forward or drop a message by taking into account the above expected utilities (Equation 3) using the softmax function as follows:

$$\begin{aligned} prob(f) &= \frac{e^{u(f)}}{e^{u(f)} + e^{u(d)}} \\ prob(d) &= \frac{e^{u(d)}}{e^{u(f)} + e^{u(d)}} \end{aligned} \tag{4}$$

Transcended Agent A transcended agent factors not just its utility and cost but also the utility of other agents with whom it identifies in its neighbourhood [5]. The notion of identifying with other agents is captured in two variables, γ (Transcendence Level), $d(i, j)$ Semantic distance between agents i and j. Expected utility of forwarding or dropping a message is computed as follows:

$$\begin{aligned} u(f) &= \frac{1}{1 + \gamma^{d(i,j)}}(-c + \gamma^{d(i,j)} * mu) \\ u(d) &= \frac{1}{1 + \gamma^{d(i,j)}}(-\gamma^{d(i,j)} * mu) \end{aligned} \tag{5}$$

Here, on forwarding a message, the intermediate agent incurs cost, mc and the sender gets a utility, mu. Since the intermediate agent identifies with the sender, it derives a scaled utility, $\gamma^{d(i,j)} * mu$. Similarly, when it drops a message, sender gets a negative utility, $-mu$ and the intermediate agent gets a scaled negative utility, $-\gamma^{d(i,j)} * mu$. The transcended agents update their semantic distances with their neighbours based on their interactions. The network settles when all transcended agents stop updating their semantic distances.

Transcended agent decides to forward or drop a message by computing probability from expected utility using a softmax function as shown in Equation 4.

Utilitarian Agent For a utilitarian agent, an action is ethical if it maximizes overall well-being. In the message passing framework, overall well-being can be accounted as overall utility. The choice of forward or drop by intermediate agent affects the utility of the sender, receiver and the agent itself. Thus, an intermediate agent driven by utilitarianism calculates the overall utility of all the stakeholders who are affected as a consequence of its action. The overall utility of forwarding or dropping a message is computed as follows:

$$\begin{aligned} u_s(f) &= mu \\ u_s(d) &= -mu \\ u_i(f) &= mu - mc \\ u_i(d) &= -mu \end{aligned} \tag{6}$$

The sender, s gets a message utility (mu or $-mu$) as a consequence of the action taken by intermediate, i. This utility is directly accounted into the expected utilities of the intermediate agent. Finally, a utilitarian agent decides to forward or drop a message by computing probabilities of forwarding and dropping calculated using the softmax function as shown in Equation 4.

RESULTS

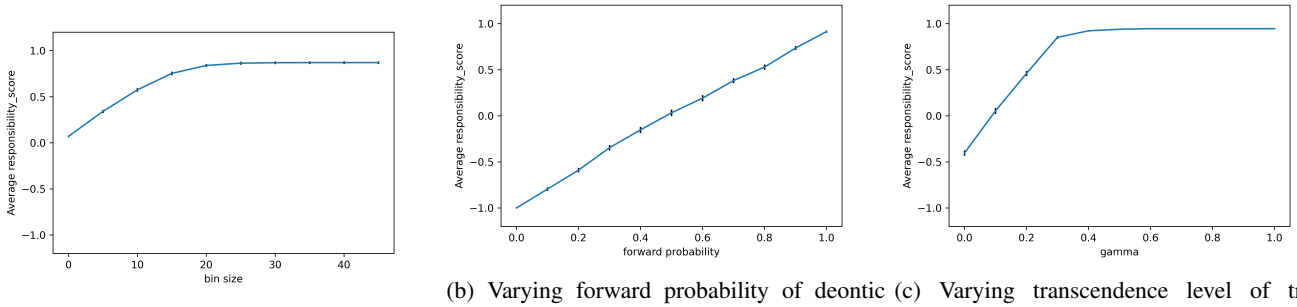
Experiments are done on an Erdős–Rényi graph with 100 nodes each representing an ethical agent. In every epoch, 1000 messages are sent in the network. Depending on the agent model, intermediate agents decide whether to forward or drop the messages. Some variants of ethical agents learn and adapt over multiple epochs. Once the system settles such that all agents reach a stable point we stop the simulation and this system state is called a stabilized network. In this stabilized network, a test epoch is simulated with another 1000 messages and the resultant metrics are recorded and presented as results. We evaluate performance of network on the following metrics:

- Expected utility: Expected utility received by source agents when their messages are forwarded by intermediate agents.
- Expected cost: Expected cost incurred by source agents in sending the messages and by intermediate agents when they forward the messages.
- Total number of burnouts: Total number of times cost incurred by agents exceeds the burnout threshold i.e. the maximum extent to which they can expend energy in forwarding the messages.
- Responsibility Score (RS): Extra number of messages forwarded f by intermediate agents, than the number of messages which were dropped d, out of all received messages. It is computed as follows for each agent:

$$RS = \frac{f - d}{f + d}$$

All the metrics are computed for each agent and then aggregated over the network. As discussed in models of ethical agents, certain hyperparameters of different ethical paradigms primarily affect the behavior of the agents. These parameters for each type of ethical agents are elaborated as follows:

- Deontic Agent: The initialized probability of agent to forward the message: initial forward probability, ifp. As deontic agents learn from their neighbourhood, initial forward probability affects how they learn about their neighbourhood and their decision to forward or drop the message.



(a) Varying bin size of a virtue ethics agent (b) Varying forward probability of deontic agent (c) Varying transcendence level of transcended agent

Fig. 3: Responsibility scores when different parameters are varied

- Virtue Agent: Maximum limit of virtue score that agent can have: bin size, bs. The motivation of agent is to be get virtue points by forwarding the message, the upper limit on that can influence its behavior.
- Transcendence Agent: Extent to which an agent identifies with other agents: transcendence level, γ. Agents take this into account when they calculate the expected utility, which further influences their decision to either forward or drop the message.
- Utilitarian Agent: Utilitarian agent maximizes the overall utility of all the stakeholders. Thus, there does not exist a hyperparameter which can influence its behavior.

Homogeneous Population

In the first part, we simulate the homogeneous population of ethical agents i.e, all the agents in the random network follow the same ethical paradigm. We varied the above mentioned hyperparameters for the respective agents and observed how these parameters influence their behavior. Results are shown in Figure 3, with average responsibility score, RS of different ethical agents in homogeneous population. The responsibility score, RS is plotted with error bars to represent the standard deviation over all agents in a network. We observe that for a deontic agent, responsible behaviour linearly increases, as the initial forward probability, ifp increases. While in case of virtue and transcended agent, responsible behaviour increases to an extent and then settles down when bin-size and transcendence level is varied respectively.

Varying burnout threshold: An agent gets burnt out when its node cost, nc exceeds the burnout threshold, bt. If an agent gets burnt out, then it can't forward a fixed number of messages and this can be understood as an agent regaining its lost energy. Using SPECTRA, we can vary the burnout threshold of ethical agents of all paradigms and analyze its resultant impact on their behavior.

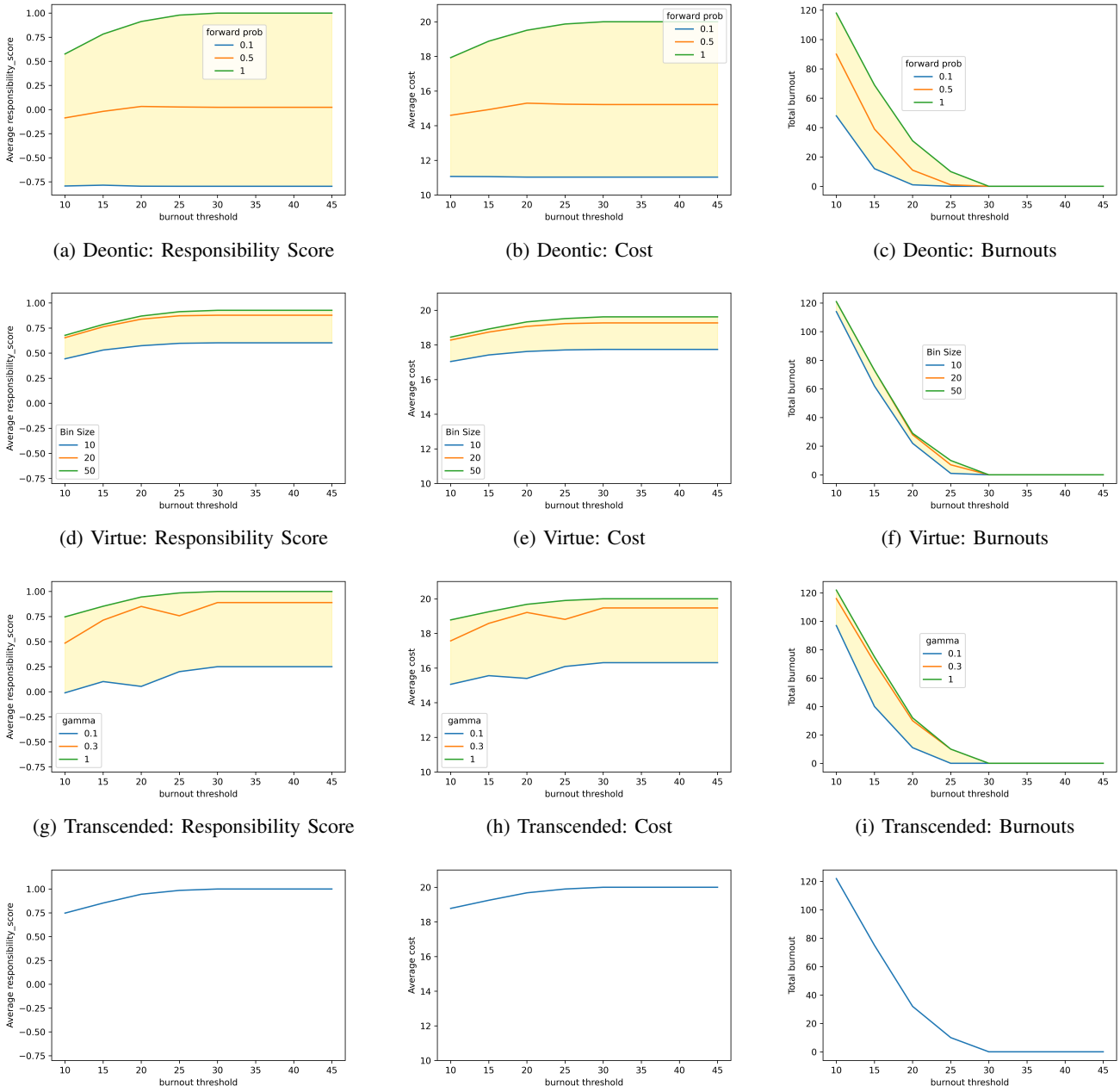
Along with varying burnout threshold, we also varied respective hyperparameters (shown by shaded region in Figure 4), to observe how hyperparameters also affects the behavior of agents while varying burnout threshold. It was observed that with increasing burnout threshold, the average responsibility score, RS of the agent also increases, as now their capacity to forward message increases. It was also observed that the average cost that agent incurs also increases, as they forward more messages. Intuitively, with increasing burnout threshold, the number of burnouts decreases. In general, it was observed that with increase in burnout threshold, all ethical agents demonstrated more responsible behavior.

In Figure 4, the shaded portion represents the spectrum of behaviors exhibited by ethical agents as their hyperparameters are varied. The shaded portion for deontic is the largest, hence varying its ifp covers a wide range of behaviors which even includes having a negative responsibility score, RS (i.e, behaving irresponsibly). While in case of virtue agent, we observe that the shaded portion in the parameter plot covers the least region when bs is varied. Hence the behavior of a virtue agent is relatively less altered by the change in its bs. In the case of transcended agent, we see an intermediate range of behavior on varying transcendence level, γ. It can be observed that even at low transcendence level, transcended agents demonstrate responsible behaviour. As mentioned earlier, utilitarian agent doesn't have a hyperparameter to be tweaked. It resembles the responsible behaviour demonstrated by transcended agent at maximum transcendence level. Hence, with the shaded portion we can infer the extent of change in behavior of ethical agents on varying model-specific hyperparameters.

Varying adversary ratio: In the next set of experiments, we introduced a proportion of adversarial agents, who forward messages with a small probability ($p = 0.05$). Using SPECTRA, we can initialize a population of agents with different behaviours, including adversarial agents. The objective of these experiments was to determine which ethical agents are sensitive in the presence of adversarial agents.

Along with varying the proportion of adversarial agents, we varied respective hyperparameters of each ethical agency. Figure 5 summarises our findings, where we plot the behavior of ethical as well as adversarial agents.

In general it was observed that with the increase in adversary ratio, the responsible behavior of deontic agents declined, while other ethical agents to a great extent were resilient against adversarial behavior. Following which the average cost and burnout rate of deontic agents see a declination as they forward lesser messages, while for the other ethical agents the trend remained almost constant. In Figure 5, the spectrum of shaded region corresponds to the range of behaviors that ethical agents exhibit while varying hyperparameters. The range of behavior with respect to ethical agent in similar as



(a) Deontic: Responsibility Score (b) Deontic: Cost (c) Deontic: Burnouts

(d) Virtue: Responsibility Score (e) Virtue: Cost (f) Virtue: Burnouts

(g) Transcended: Responsibility Score (h) Transcended: Cost (i) Transcended: Burnouts

(j) Utilitarian: Responsibility Score (k) Utilitarian: Cost (l) Utilitarian: Burnouts

Fig. 4: Varying burnout threshold for different types of ethical agents

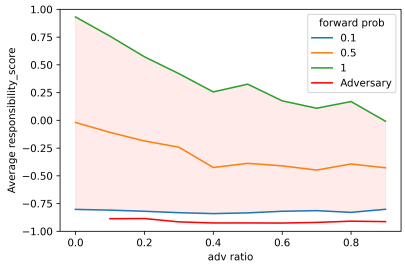
we discussed the case of varying burnout threshold. The trends for adversarial agents were almost constant, as their behaviour is unaffected by their neighbourhoods.

Mixed Population

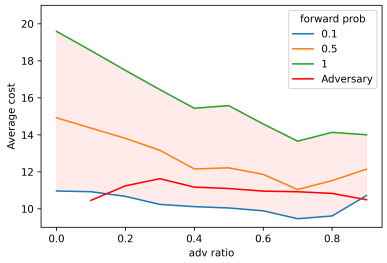
We have looked at homogeneous populations of agency following a specific ethical model of responsible behaviour in context of a message passing network. However, a more realistic setup is where agents with different ethical models interact with each other. These experiments and results will be useful to understand how the interactions between agents of different models of ethics, impact individual agents, and the system as a whole.

For this experiment, we consider a network of 400 agents with equal proportion of each ethical paradigm. In a network, factors like degree of a node, position of the node (for instance leaf node) etc. can confound the metrics being measured. Thus, we handle the confounding effect of network topology on each ethical paradigm by using stochastic averaging. We run 1000 simulations with random initialization of the nodes, and present results of averaged metrics over the runs with the error bars representing standard deviation.

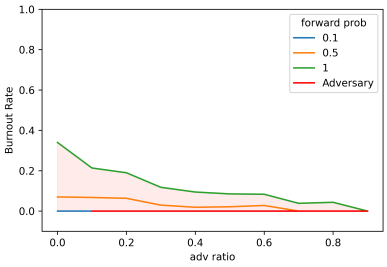
In the previous section, agents only interacted with adver-



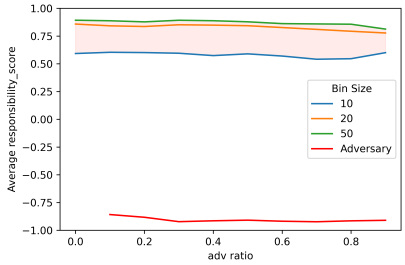
(a) Deontic: Responsibility Score



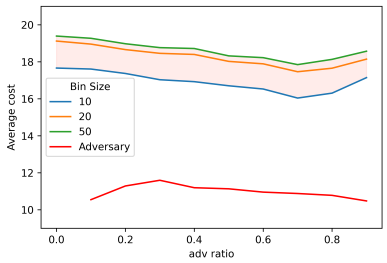
(b) Deontic: Cost



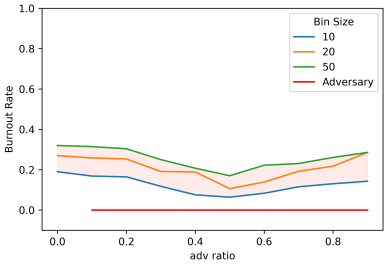
(c) Deontic: Burnout Rate



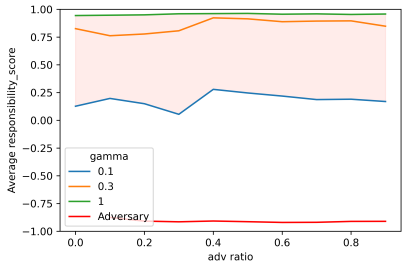
(d) Virtue: Responsibility Score



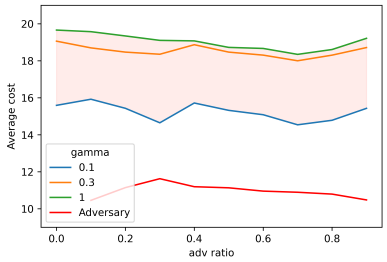
(e) Virtue: Cost



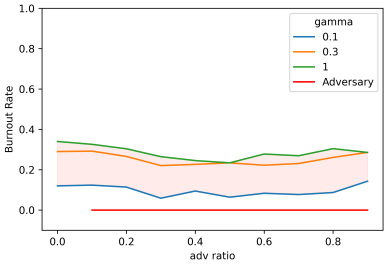
(f) Virtue: Burnout Rate



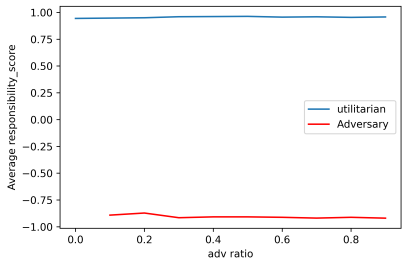
(g) Transcended: Responsibility Score



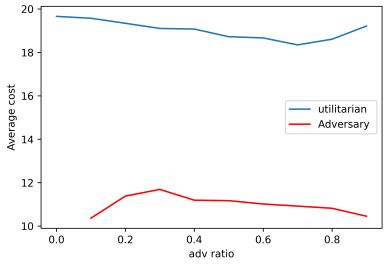
(h) Transcended: Cost



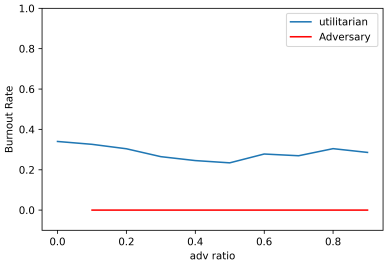
(i) Transcended: Burnout Rate



(j) Utilitarian: Responsibility Score

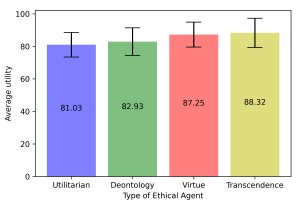


(k) Utilitarian: Cost

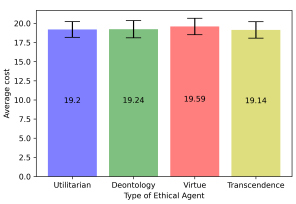


(l) Utilitarian: Burnout Rate

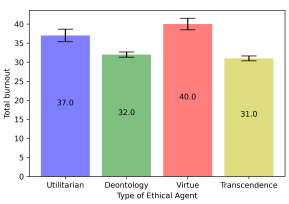
Fig. 5: Varying adversary ratio for different types of ethical agents



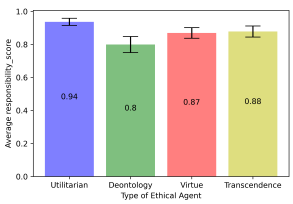
(a) Utility



(b) Cost



(c) Burnouts



(d) Responsibility score

Fig. 6: Results for mixed population of ethical agents

sarial agents apart from agents of their own type. In this case, they interact with ethical agents following different models of ethics. We note that the responsibility score for the utilitarian agents is highest since they forward the most number of messages. However, they also burnout a lot. They get the lowest utility which denotes that their messages are forwarded the least, despite their best behaviour. Since deontic agents learn and adapt to their neighbourhood, they are sensitive to irresponsible behavior. Thus, their responsibility score is lowest, and their burnouts are low. Virtue agents regulate their forwards and drops around a threshold and in this process they get burnt-out the most. Finally, transcended agents have highest utility and the lowest number of burnouts, while demonstrating high responsible behaviour. These trends might change if the proportion of different types of ethical agents in the network is varied.

DISCUSSION

We modeled ethical agents following a variety of models of ethics, namely– utilitarianism, deontology, virtue ethics and transcendence. Utilitarian agents maximize the utility of the collective. Deontic agents adopt the network's notion of ethics. They seek to affiliate with other agents in the network and conform with the behaviour of their neighbourhood. Virtuous agents focus on demonstrating a context-specific virtue. And transcended agents have an elastic sense of self such that they identify with other agents in their neighbourhood.

These models of ethics are not completely independent to each other. We observe some commonalities across different models of ethics, which are discussed as follows. Highest transcendence level looks similar to utilitarianism, as agents at maximum transcendence level ($\gamma = 1$) account for other agents to the maximum extent. However, utilitarian agents always consider all stakeholders equally while transcended agents have the capability to adapt based on their interactions with individual stakeholders. Transcendence trends also look similar to virtue ethics (as shown in Figures 3a and 3c). While transcended agents on increasing transcendence level demonstrate maximum responsibility score, virtue agents settle at a lower responsibility score. Virtue agents only focus on demonstrating virtuous behaviour whereas transcended agents demonstrate virtuous behaviour while also fulfilling their self-interest. Also, in case of transcendence, responsible behaviour is an emergent characteristic rather than something which agents are forced to uphold. Transcendence gives flexibility to agents to adapt to other agents and the environment based on changing context.

CONCLUSIONS AND FUTURE WORK

In this paper, we introduce SPECTRA which provides a common platform to quantitatively compare different models of computational ethics. While different models of ethics have been evaluated in different contexts, to the best of our knowledge, there is no common evaluation test-bed across these models. SPECTRA also enables the system designer to analyze fine-grained differences between different ethical theories which help in making informed decisions about which paradigm to use in which setting.

The test-bed can be extended in a variety of ways to incorporate different variations of AAs and environment. Currently, the agents only account for their $1-$hop neighbours. In future, it can be extended to consider agents that are multiple hops away. The core dilemma can be further extended to include other ethical implications like collateral effects on indirectly affected entities in the system.

ACKNOWLEDGMENT

We thank Machine Intelligence and Robotics (MINRO) center funded by Government of Karnataka, India and Centre for Internet of Ethical Things (CIET) funded by Government of Karnataka, India and World Economic Forum for supporting this work.

REFERENCES

[1] M. Pechoucek, S. G. Thompson, and H. Voos, *Defence Industry Applications of Autonomous Agents and Multi-Agent Systems*. Springer, 2008.

[2] M. Pěchouček and V. Mařík, "Industrial deployment of multi-agent technologies: review and selected case studies," *Autonomous agents and multi-agent systems*, vol. 17, no. 3, pp. 397–431, 2008.

[3] A. F. Winfield and M. Jirotka, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180085, 2018.

[4] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein, "Implementations in machine ethics: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 6, pp. 1–38, 2020.

[5] J. Deshmukh and S. Srinivasa, "Computational transcendence: Responsibility and agency," *Frontiers in Robotics and AI*, vol. 9, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2022.977303

[6] C. Allen, I. Smit, and W. Wallach, "Artificial morality: Top-down, bottom-up, and hybrid approaches," *Ethics and information technology*, vol. 7, no. 3, pp. 149–155, 2005.

[7] D. Abel, J. MacGlashan, and M. L. Littman, "Reinforcement learning as a framework for ethical decision making," in *Workshops at the thirtieth AAAI conference on artificial intelligence*, 2016.

[8] C. Cloos, "The utilibot project: An autonomous mobile robot based on utilitarianism," in *2005 AAAI Fall Symposium on Machine Ethics*, 2005, pp. 38–45.

[9] M. Anderson, S. L. Anderson, and C. Armen, "Towards machine ethics," in *AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA*, 2004.

[10] C. Van Dang, T. T. Tran, K.-J. Gil, Y.-B. Shin, J.-W. Choi, G.-S. Park, and J.-W. Kim, "Application of soar cognitive agent based on utilitarian ethics theory for home service robots," in *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*. IEEE, 2017, pp. 155–158.

[11] L. A. Dennis, M. Fisher, and A. Winfield, "Towards verifiably ethical robot behaviour," in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[12] D. Vanderelst and A. Winfield, "An architecture for ethical robots inspired by the simulation theory of cognition," *Cognitive Systems Research*, vol. 48, pp. 56–66, 2018.

[13] A. F. Winfield, C. Blum, and W. Liu, "Towards an ethical robot: internal models, consequences and ethical action selection," in *Conference towards autonomous robotic systems*. Springer, 2014, pp. 85–96.

[14] I. Kant and J. B. Schneewind, *Groundwork for the Metaphysics of Morals*. Yale University Press, 2002.