Evaluating Computational Models of Ethics for Autonomous Decision Making

Janvi Chhabra[1*†], Karthik Sama[1†], Jayati Deshmukh[1*], Srinath Srinivasa[1]

[1*]Department, International Institute of Information Technology, Bangalore, 26/C, Electronics City, Hosur Road,, Bangalore, 560100, Karnataka, India.

*Corresponding author(s). E-mail(s): janvi.chhabra@iiitb.ac.in; jayati.deshmukh@iiitb.org;
Contributing authors: sai.karthik@iiitb.ac.in; sri@iiitb.ac.in;
[†]These authors contributed equally to this work.

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 have 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 emergent properties of the system.

Keywords: Ethics, Responsible Agency, Agent-based Modelling, Multi-Agent Systems

1 Introduction

Autonomous agents are becoming prevalent in current times across a variety of application areas like autonomous vehicles, autonomous industrial bots, autonomous weapon systems etc. [36, 35]. The underlying algorithms of these autonomous agents are designed such that they have a high level of autonomy and can operate with minimal or no external feedback. They mostly operate in systems consisting of other agents as well as humans [30] and therefore their decisions and actions directly affect others. Thus, autonomous agents must consider ethical aspects before taking any action [17]. Ethical autonomous agents have higher acceptance in societies since they can be trusted to act aligned to specific paradigms of ethics [47].

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 argument for the way ethical dilemmas are approached. In realistic situations, designing ethical autonomous agents may require some of these paradigms of ethics, or a combination of these to act as an underlying foundation for the application. Also, games and simulations are useful techniques to evaluate the morality of virtuous agents [44, 18].

Broadly, paradigms of ethics can be classified into three overarching classes: *utilitarianism*, *deontological ethics* and *virtue ethics* [40, 16]. *Utilitarianism* resolves ethical dilemmas based on the 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 the 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* [15] argues ethical choices are a natural consequence of agents *identifying* themselves with other agents. This is modelled using an *elastic sense of self* in an agent such that its perceived utility is a function of not only its own payoffs but also payoffs accrued by all stakeholders that it identifies with. With an elastic sense of self, responsible behaviour is shown to be a *natural consequence* of self-interest dynamics, rather than something that conflicts with self-interest.

Autonomous agents can operate on either of these paradigms of ethics and as discussed above, each paradigm leads to an action due to an underlying argument. How do ethical agents behave in case of responsibility dilemma of choosing between individual benefit and collective good? How do autonomous agents decide which paradigm to operate on? What happens in a scenario where multiple autonomous agents are 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 kinds of questions, we need a testbed that can simulate and test different paradigms of ethics and in turn compare and contrast them.

In this paper, we present SPECTRA - Strategic Protocol Evaluation and Configuration Testbed for Responsible Autonomy which is designed as a framework to compare autonomous agents with different models of ethics. It is built on a network representing a multi-agent setup, where different ethical agents interact with each other over time. The framework presents a "responsibility dilemma" i.e. a scenario with conflicting choices, one which is individually beneficial for the agent and the other which is individually suboptimal for the agent but good for the system or the collective. "Responsibility" can be modelled in multiple ways [42], however, in the context of this paper, we call autonomous agents to be responsible if they can act ethically accounting for the impact of their actions on the system or the collective and can resolve the responsibility dilemma in the context in which they are operating. Agents based on their model of ethics try to resolve this responsibility dilemma and demonstrate ethical (or unethical) behaviour in different ways. We then use the outputs of the SPECTRA framework to analyze the impact of the agents' behaviour on individual agents and the system as a whole.

The paper is organized as follows: In Section 2, we introduce the paradigms of ethics which can be used as a model of ethics in autonomous agents. Especially we elaborate on utilitarianism, deontology, virtue ethics and computational transcendence. Next, in Section 3, we present the SPECTRA framework and demonstrate how it can be used to model different kinds of ethical agents in a multi-agent network scenario. We discuss the specifics like the required hyperparameters and algorithmic details of all models of ethics using the SPECTRA framework. In Section 4, we analyze the results of a homogenous population of ethical agents i.e. where all the agents in the network have the same model of ethics and of a heterogenous population of ethical agents i.e. where different agents in the network are aligned with different models of ethics. While we present simulation results for the four paradigms of ethics, the testbed is generic and can be used to simulate other paradigms of ethics as well. We discuss the benefits and novelty of the SPECTRA framework and how it is useful to compare different models of ethical autonomous agents in Section 5. Finally, we conclude in Section 6 where we highlight some of the key takeaways of this work and also highlight various ways in which it can be extended in future.

2 Ethical Paradigms

In this section, we discuss the following well-known ethical paradigms: *utilitarianism*, *deontological ethics* and *virtue ethics* [40, 16]. We also introduce *Computational Transcendence* [15], a recently proposed ethical model of agency. An understanding of these ethical schools of thought will be useful in building computational models of different ethical agents.

2.1 Utilitarianism

Utilitarianism or consequential ethics [40, 2, 32, 31] 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 is 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 [1, 9, 5, 41, 12, 43, 46, 5].

Some of the challenges of consequentialism include difficulty in evaluating consequences, especially in open-world conditions with uncertainty. Defining utility–

especially long-term notions of “greater good” is yet another challenge. In extensive games, computing long-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.

2.2 Deontology

Deontological ethics [40] 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” [23]. For example, one should not lie because if everyone lies, then no one will trust anyone and human communication will lose its value. Kantian cooperation is a way to design cooperative agents where agents decide a strategy which they would prefer all agents to choose [24].

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

A variety of computational models of ethics have been designed using the deontological paradigm. Some examples include BDI (Beliefs, Desires and Intentions) and its variants [11, 33], inductive logic and first-order logic [4, 29], normative models like OPF (obligated, permitted and forbidden) constraints [25, 45] and rule-learning based models [27, 34]. There are models based on utility calculus that use deontological statements as a prior for utilitarian model [39]. Some of these models have also been applied in use-cases like medical AI [6], robots for elder care and for assisting patients [3, 38] and even war-fighting robots [8, 37].

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 that can occur in the system must be known in advance so that they can be handled appropriately. Conflicting rules are yet another challenge with deontological paradigms– especially in large state spaces with multiple considerations.

2.3 Virtue Ethics

Virtue ethics [40] 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 focus on either actions or consequences, but on agents themselves and if they are displaying virtuous behaviour.

Aristotle elaborated on some of the core ideas of virtue ethics. He believed that to “flourish”, 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.” In simple words, according to the role an agent has to perform, the agent should try to attain the virtues that 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 virtues is a continuously ongoing process, thus agents should practice and demonstrate virtuous behaviour over time.

Being a virtuous agent is not just about the attainment of virtues, but also about attaining the "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 practising, an agent figures out what works best in that context.

Some computational implementations of virtue ethics are as follows: Agents can learn virtues based on specific contexts using artificial neural networks [22, 19, 20, 21]. TruthTeller and Sirocco can learn and reason from moral data with the purpose of supporting humans in ethical reasoning [26, 28]. Imitation learning through the role of moral exemplars [7, 18]. Goal-directed virtuous agents whose virtues should be defined based on the role or function they are performing. Also, autonomous agents need not imbibe human virtues rather they should demonstrate virtues that are relevant for machines in their context [10].

2.4 Computational Transcendence

Computational transcendence [15] approaches the question of ethics by imbibing agents with an *elastic sense of self* that enables them to *identify* with other agents and groups of agents. In the case of transcended agents, responsible choice for collective welfare is the natural consequence of them identifying themselves with other agents. Specifically, transcended agents have two main parameters using which they adjust identification with others– transcendence level or elasticity γ, denotes the extent to which these agents are open to identifying with some external entity, and semantic distance d denoting an attenuation rate for the agent's sense of identity with each object they identify with.

Rational associations between agents only last as long as the association serves the self-interest of the agents involved; in contrast, in associations of identity, agents identifying with one another act as if the other agent's interests were one's own interest and strategize accordingly. Transcended agents derive utility from their payoffs as well as the payoffs of the external entities that they identify with. Computational transcendence has been demonstrated to be effective in various real-world applications like supply chains and traffic management [13, 14].

Computational transcendence presents a paradigmatic departure in this setting. Virtuous or cooperative behaviour *emerges* as a result of the transcendence of their sense of self. This transcendence, too, is regulated by rational rather than normative considerations– a transcended agent can adjust its transcendence level and/or semantic distance to adapt based on how generous and trustworthy its environment is. Utilitarianism requires the utility of all agents in the system to be computed and made available for each agent as common knowledge. In contrast, with computational

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

transcendence, responsible behaviour can emerge with local knowledge. However, computational transcendence still requires this local knowledge– in the form of knowledge of payoffs of other agents with which it is interacting, in order to compute one's transcended utility.

3 SPECTRA Framework

To be able to compare and contrast the different paradigms of responsible agency, we will need a common testbed. Towards this end, we present *SPECTRA* (Strategic Protocol Evaluation and Configuration Testbed for Responsible Autonomy) which is a multi-agent testbed 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 versus 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 versus responsible choice.

We model the interactions among agents using an undirected graph, where a node represents an autonomous agent and the presence of an edge between two agents represents that those agents can interact. In an interaction, an agent can assume the 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 a message from the sender and decides whether to forward the message to the receiver or to drop the message.
(c) Receiver: Receives the message sent by the sender via an intermediate agent.

The matrices of node utility, nu and node cost, nc for the sender, s and intermediate, i are shown in Table 1. 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.

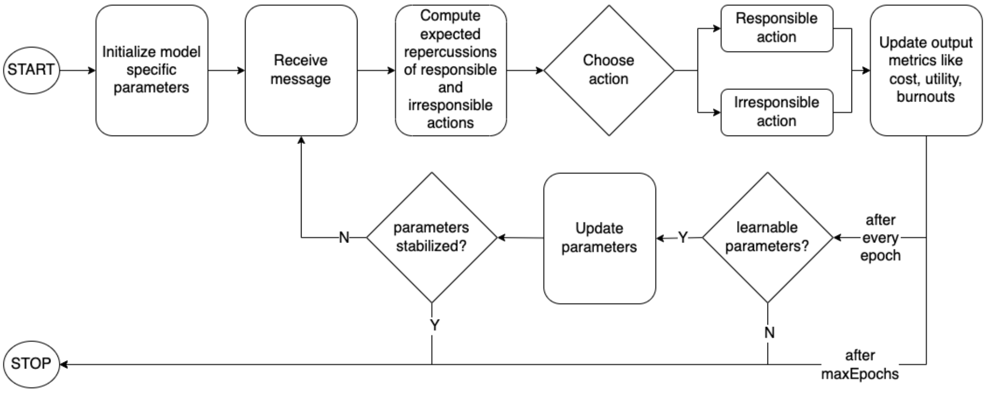


Fig. 1: Block diagram of an Ethical Agent

Thus, there is no rational incentive for the intermediate agents to forward messages. However, if intermediate agents don't forward messages, then no messages would reach the intended receivers resulting in a network that 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 the 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 the 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 that make choices taking collective welfare into consideration instead of just their self-interest. The testbed 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 autonomous agents following different paradigms of ethics in a network and study the impact on the system as a whole.

3.1 Computational Models of ethical agents

Using the SPECTRA framework, we model ethical agents driven by different paradigms of ethics like deontology, virtue ethics and utilitarianism. We also model transcended agent which operates based on a recently proposed model called Computational Transcendence [15]. Agents decide to forward or drop a message based on the ethical paradigm they follow. Qualitatively, every agent differs in the following aspects:

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

The overall block diagram of an ethical agent is shown in Figure 1 and a comparison between different key ethical paradigms is shown in Table 2. We now elaborate on how theoretical constructs for responsible behaviour are modelled in agents following

	Deontology	Virtue Ethics	Transcendence	Utilitarianism
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
Computational Approach	Learn the norms from its neighbourhood.	Gain virtue points by performing virtuous actions.	Transcend to others and include their interests.	Maximize overall utility for all stakeholders.
Hyperparameter	Initial forward probability	Bin size	Transcendence Level	None

Table 2: Comparison between different models of ethics

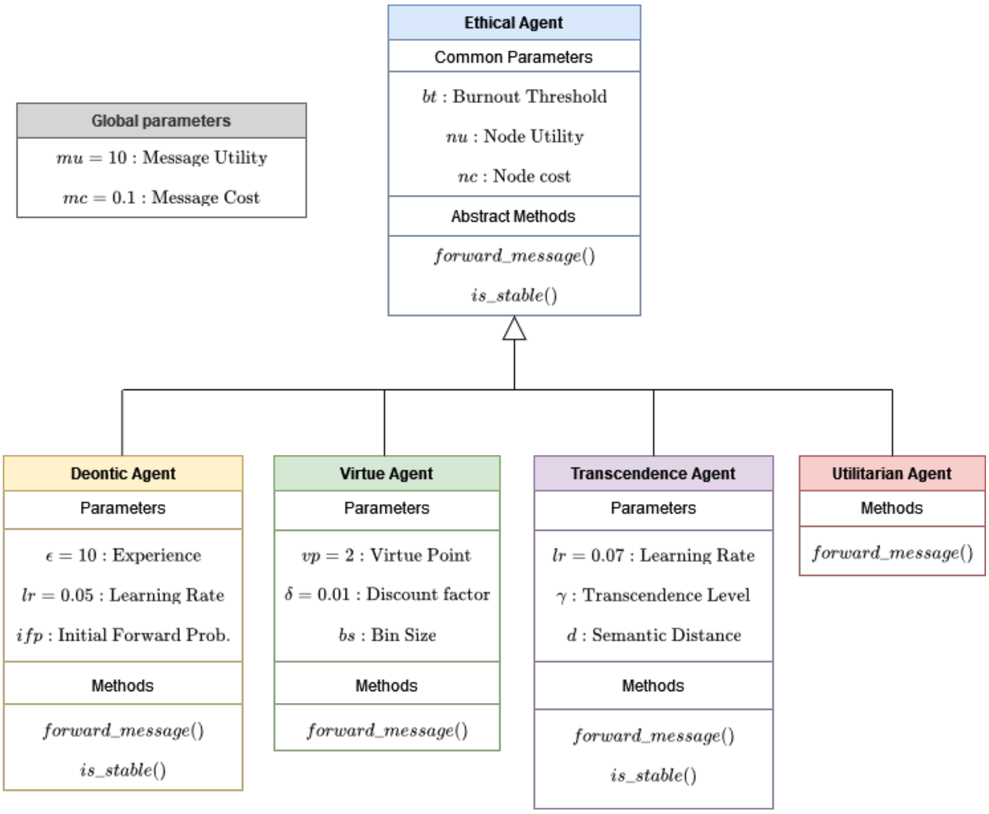


Fig. 2: Class diagram for the SPECTRA Framework

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 ethical agent class is an abstract class that can be inherited by a child class to implement a specific model of ethics. The $forward\_message()$ and $is\_stable()$ are the abstract methods in the ethical agent class. The utility obtained by the sender upon its message being delivered to the recipient is mu (Message Utility). While the cost an agent incurs for transmitting a message is mc (Message Cost). Both mu and mc are global parameters. The common parameters, nu (Node Utility) and nc (Node Cost) are the total utility and total cost accrued by an agent respectively.

System designers can model their variants of ethical agents by implementing abstract methods as per their interpretation of ethics and ethical paradigms in the given context. We present the implementation of four such ethical agent classes and justify how the ideas of some of the paradigms of ethics in Section 2 translate into corresponding computational models in SPECTRA.

Algorithm 1: Deontic agent

```
Function forward_message():
    if $messagesSent < \epsilon$ then
        $fp_t = ifp$
    else
        $rr = \frac{\text{messagesReached}}{\text{messagesSent}}$
        $fp_t = (1 - lr) * fp_{t-1} + (lr * rr)$
    if $random() < fp_t$ then
        return True
    else
        return False
End Function
Function is_stable():
    if $fp_t - fp_{t-1} < fp_{threshold}$ then
        return True
    else
        return False
End Function
```

3.1.1 Deontic Agent

A deontic agent operates on Immanuel Kant's "Categorical Imperative" which has been discussed in Section 2.2. 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 actions with them.

To model this in agents, we introduce 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 settled message-passing network. In a population of diverse ethical agents, the significance of the hyperparameter ifp reduces for a deontic agent, as it starts learning from agents driven by other ethical principles.

We define rr (Reach Ratio) as the ratio of messages of the agent that have reached their destination over all the messages sent by the agent. The implementation of $forward\_message()$ and $is\_stable()$ functions for deontic agent is described in Algorithm 1.

Algorithm 2: Virtue agent

```
Function forward_message():
    $c = \delta * (nc - nu) + mc$
    $utility\_forward = min(vp, bs - vs) - c$
    $utility\_drop = -vp$
    $prob\_forward = softmax(utility\_forward)$
    $prob\_drop = softmax(utility\_drop)$
    if $prob\_forward > prob\_drop$ then
        return True
    else
        return False
End Function
Function is_stable():
    return True
End Function
```

3.1.2 Virtue Agent

The virtue ethics model neither focuses on the action nor its consequences but on the demonstration of virtuous behaviour by the agent. In the message passing scenario, we have modelled 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 introduce 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 Section 2.3, 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 constitutes these extreme behaviours.

The idea of the golden mean has been modelled in a virtue agent by placing an upper bound on the vp that it can accumulate and taking into account the cost it incurs upon forwarding a message. 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 a hyperparameter which will be referred to as bs. At any instant, the aggregate of vp in this bin is considered vs (Virtue Score) of the agent. The virtue agent computes the overall cost c incurred for forwarding a message, which consists of a relative historic cost, scaled down by δ, and the current message forwarding cost mc. This helps it to maintain a balance between the cost it incurs and the virtue points it accumulates.

Virtue Score vs of an agent can't exceed the bin size bs. On forwarding a message, the agent accumulates vp up to the limit of bs. The utility that the virtue agent gets on forwarding is the net difference between the virtue points it accumulates and the cost it incurs. On dropping the message, vp is deducted from the virtue score, the agent gets utility of $-vp$. A virtue agent computes the utility it gets by forwarding or dropping a message as $utility\_forward$ and $utility\_drop$ respectively. Finally, the virtue agent calculates the probability of forwarding or dropping a message by passing these utilities into the softmax function. In the case of a virtue agent, it doesn't learn any parameter, hence the $is\_stable()$ method always returns true. The implementation of $forward\_message()$ and $is\_stable()$ functions for virtue agent is described in Algorithm 2.

3.1.3 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 [15]. The notion of identifying with other agents is captured in two variables, γ (Transcendence Level) and $d(i,j)$ Semantic distance between agents i and j. The expected utility of forwarding or dropping a message is denoted as $utility\_forward$ and $utility\_drop$.

On forwarding a message, the intermediate agent incurs a 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, the 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 stabilizes when all transcended agents stop updating their semantic distances. The transcended agent decides to forward or drop a message by computing probability from their respective expected utilities using a softmax function. The implementation of $forward\_message()$ and $is\_stable()$ functions for transcended agent is described in Algorithm 3.

3.1.4 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. Thus, an intermediate agent driven by utilitarianism calculates the overall utility of all the stakeholders who are affected as a consequence of its action. A utilitarian agent computes the utility it gets by forwarding or dropping a message as $utility\_forward$

Algorithm 3: Transcended agent

```
Function forward_message(source)():
    $cost = \delta * (nodeCost - nodeUtility) + messageCost$
    $d = distance(intermediate, source)$
    $utility\_forward = \frac{-cost + \gamma^d * messageUtility}{1 + \gamma^d}$
    $utility\_drop = \gamma^d * messageUtility$
    $prob\_forward = softmax(utility\_forward)$
    $prob\_drop = softmax(utility\_drop)$
    if $prob\_forward > prob\_drop$ then
        return True
    else
        return False
End Function
Function is_stable():
    for $(node_i, node_j)$ in network do
        $d_t = distance_t(node_i, node_j)$
        $d_{t-1} = distance_{t-1}(node_i, node_j)$
        if $|d_t - d_{t-1}| > distance_{threshold}$ then
            return False
    end
    return True
End Function
```

and *utility_drop*, respectively. Further, it will calculate the probability of forwarding or dropping the message by passing these utilities into a softmax function. As there are no learnable parameters, the *is_stable*() method for a utilitarian agent always returns true. The implementation of *forward_message*() and *is_stable*() functions for utilitarian agent is described in Algorithm 4.

The code for the SPECTRA framework is available at https://github.com/WSL-IIITB/Computational-Ethics. One can extend the framework for inculcating other models of ethical paradigms.

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 ethical paradigm followed by the agent, it decides 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 the performance of the network on the following metrics:

Algorithm 4: Utilitarian agent

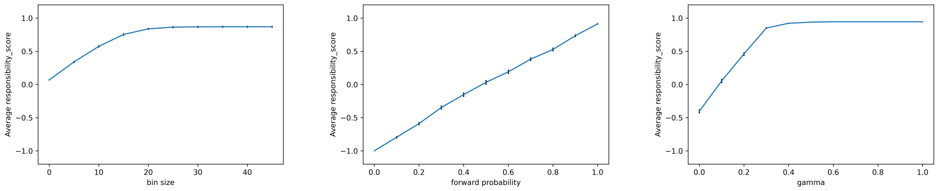
```
Function forward_message():
    $utility\_forward = mu - mc$
    $utility\_drop = -mu$
    $prob\_forward = softmax(utility\_forward)$
    $prob\_drop = softmax(utility\_drop)$
    if $prob\_forward > prob\_drop$ then
        return True
    else
        return False
End Function
Function is_stable():
    return True
End Function
```

- 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 Section 3.1, certain hyperparameters of different ethical paradigms primarily affect the behaviour of the agents. These parameters for each type of ethical agent are elaborated as follows:

- Deontic Agent: The probability with which deontic agent initially forwards a message, initial forward probability ifp. It affects what a deontic agent learns from its neighbourhood and subsequently, its decision to forward or drop a message.
- Virtue Agent: Maximum limit of virtue score that the agent can accrue, bin size bs. Virtue points motivate a virtue agent to forward a message. However, an upper limit on virtue score regulates the extent to which, the reliability virtue gets expressed by the agent.
- 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.



(a) Varying the 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 hyperparameters are varied

- Utilitarian Agent: Utilitarian agent maximizes the overall utility of all the stakeholders. Thus, there does not exist a hyperparameter which can influence its behaviour.

4.1 Homogeneous Population

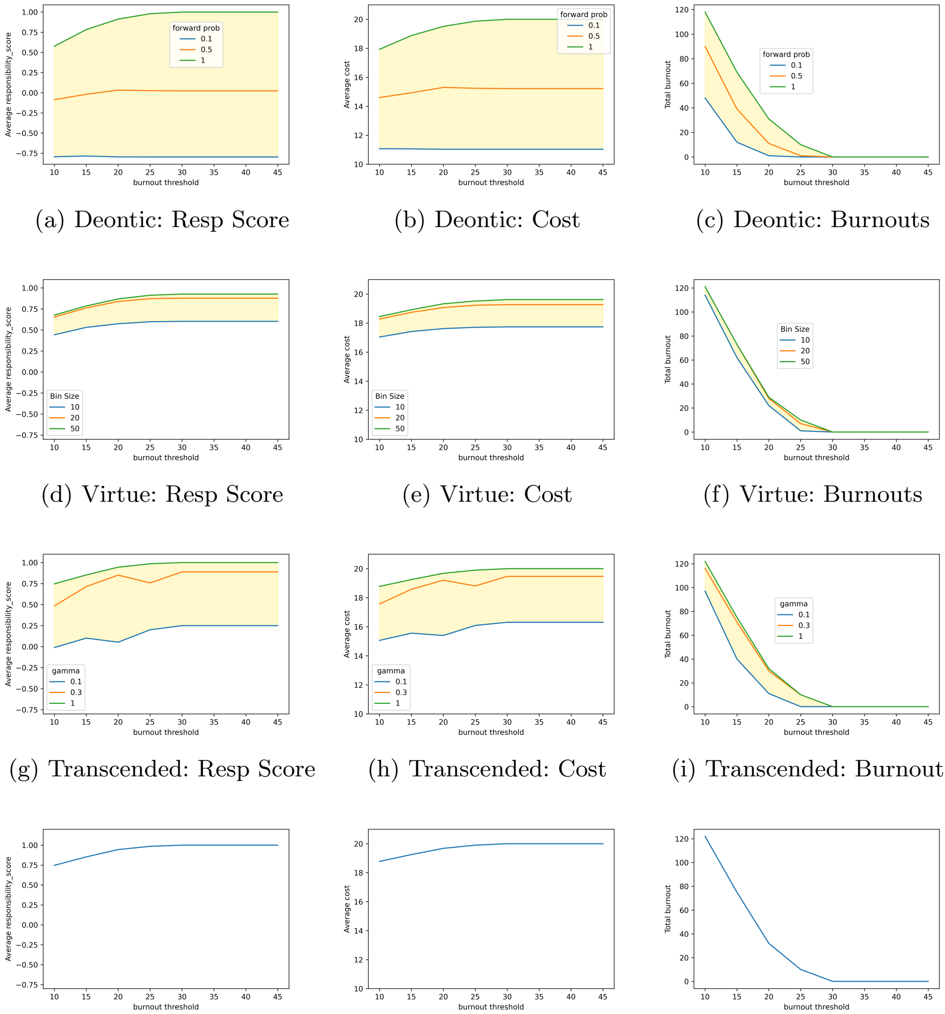
In the first part, we simulate a homogeneous population of ethical agents i.e., all the agents in the network follow the same ethical paradigm. We vary the above mentioned hyperparameters for the respective agents and observe how they influence their behaviour. Plots of average responsibility score, rs of different ethical agents in homogeneous populations are shown in Figure 3. It 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 increases. In the case of virtue and transcended agent, responsible behaviour increases to an extent and then settles down when bin size and transcendence level are varied respectively. Though virtue agents and transcended agents show similar trends, we observe that the responsibility score for a transcended agent settles at a higher value than the virtue agent.

4.1.1 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 while it regains its lost energy. Using SPECTRA, we can vary the burnout threshold of ethical agents of all paradigms and analyze its resultant impact on their behaviour.

We observe that with increasing burnout threshold, the average responsibility score, rs of the ethical agents also increases, as their capacity to forward messages increases. Also, the average cost incurred by agents also increases, as they forward more messages. Intuitively, with increasing burnout threshold, the number of burnouts decreases. In general, we note that with an increase in the burnout threshold, all ethical agents demonstrate more responsible behaviour.

We also varied respective hyperparameters (shown by the shaded region in Figure 4), to observe how it affects the behaviour of agents while varying the burnout



(a) Deontic: Resp Score

(b) Deontic: Cost

(c) Deontic: Burnouts

(d) Virtue: Resp Score

(e) Virtue: Cost

(f) Virtue: Burnouts

(g) Transcended: Resp Score

(h) Transcended: Cost

(i) Transcended: Burnouts

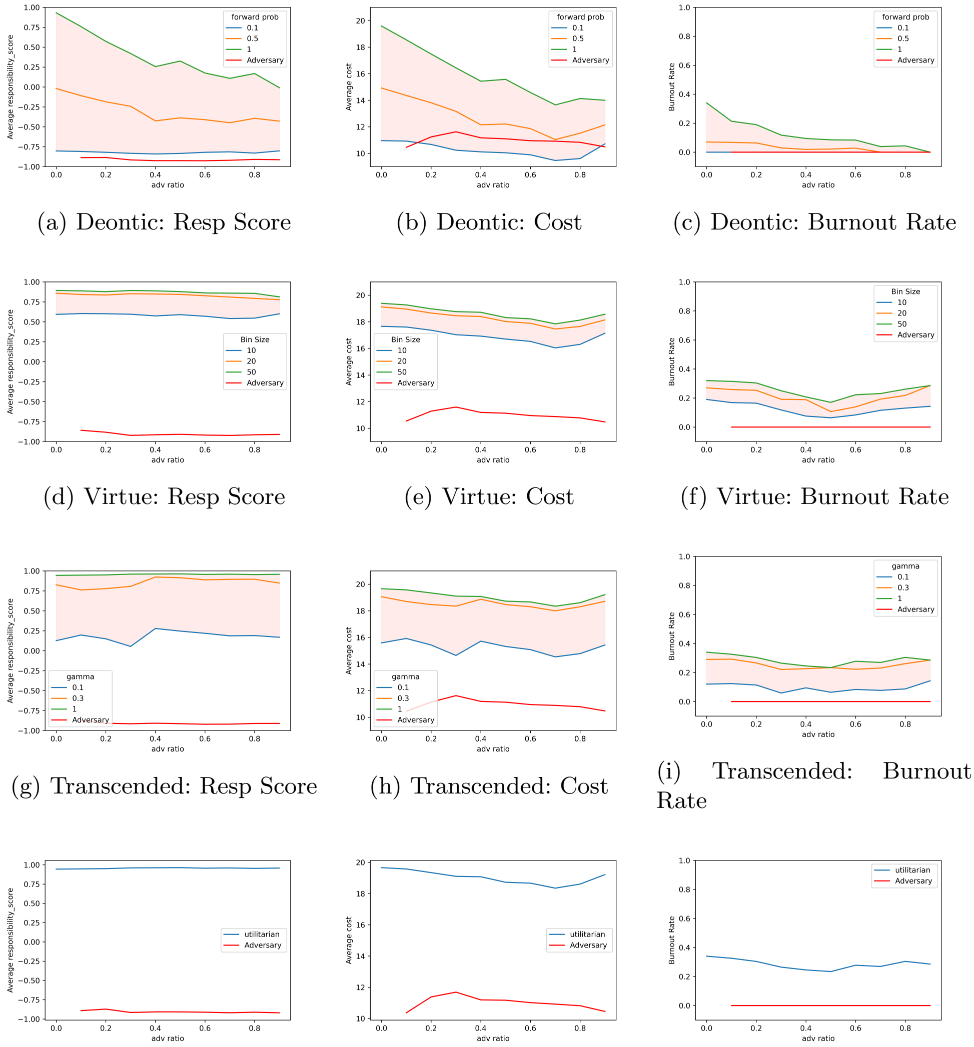
(j) Utilitarian: Resp Score

(k) Utilitarian: Cost

(l) Utilitarian: Burnouts

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

threshold. The shaded portion for deontic is the largest, hence varying its ifp covers a wide range of behaviours which even includes having a negative responsibility score, rs (i.e., behaving irresponsibly). In the case of the virtue agent, we observe that the shaded portion in the parameter plot covers the least region when bs is varied. Hence the behaviour of a virtue agent is relatively less altered by the change in its bs. In the case of a transcended agent, we see an intermediate range of behaviour on varying transcendence level, γ. It can be observed that even at a low transcendence level, transcended agents demonstrate responsible behaviour. As mentioned earlier,



(a) Deontic: Resp Score

(b) Deontic: Cost

(c) Deontic: Burnout Rate

(d) Virtue: Resp Score

(e) Virtue: Cost

(f) Virtue: Burnout Rate

(g) Transcended: Resp Score

(h) Transcended: Cost

(i) Transcended: Burnout Rate

(j) Utilitarian: Resp Score

(k) Utilitarian: Cost

(l) Utilitarian: Burnout Rate

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

a utilitarian agent doesn't have a hyperparameter to be tweaked. It resembles the responsible behaviour demonstrated by the transcended agent at the maximum transcendence level. Hence, with the shaded portion, we can infer the extent of change in the behaviour of ethical agents on varying model-specific hyperparameters.

4.1.2 Varying adversary ratio

In the next set of experiments, we introduce 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 is to determine which kind of ethical agents are sensitive to the presence of adversarial agents.

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

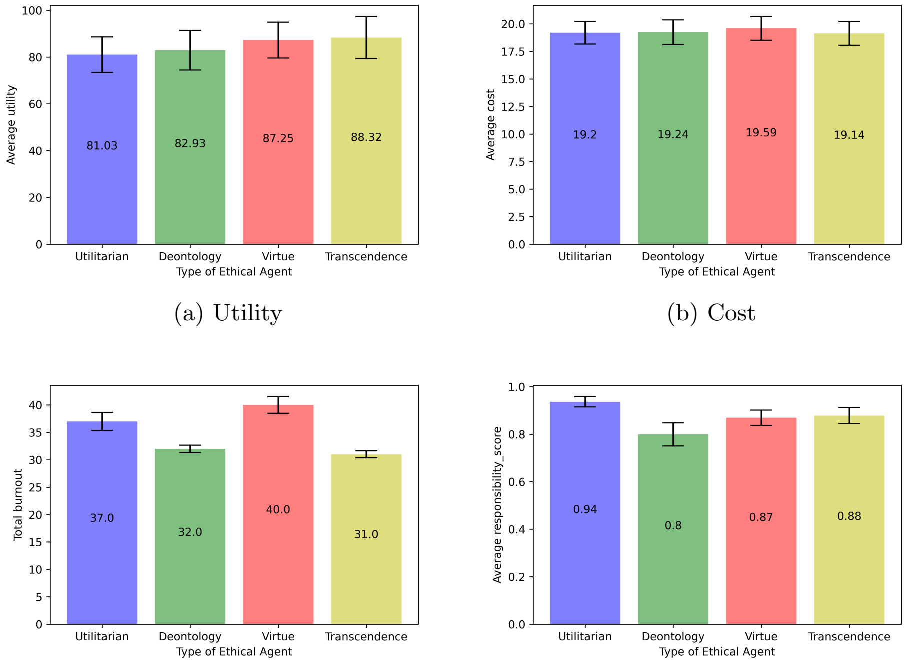
In general, it is observed that with the increase in adversary ratio, deontic agents are most sensitive to adversaries, and they demonstrate irresponsible behaviour, while virtue agents are the most resilient against adversaries in the spectrum of behaviours exhibited. Due to this, the average cost and burnout rate of deontic agents declines as they forward fewer messages, while for the other ethical agents, the trend remains almost constant. In Figure 5, the spectrum of the shaded region corresponds to the range of behaviours that ethical agents exhibit while varying hyperparameters. The range of behaviour with respect to ethical agents is similar as discussed in the case of varying burnout threshold. The trends for adversarial agents are almost constant, as their behaviour is unaffected by their neighbourhoods.

4.2 Mixed Population

In section 4.1, we looked at homogeneous populations of agents following a specific ethical paradigm in the context of a message-passing network. However, a more realistic setup is where agents with different ethical paradigms interact with each other. These experiments and results are useful for understanding the impact of interactions between agents of different models of ethics on individual agents and the system as a whole.

For this experiment, we consider a network of 400 agents with an equal proportion of each ethical paradigm. In a network, factors like the 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 Section 4.1.2, agents only interacted with adversarial agents apart from agents of their own type. In this case, they interact with ethical agents following different models of ethics. The simulation results are shown in Figure 6. 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 neighborhood, they are sensitive to irresponsible behaviour. 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 (with $\gamma = 0.25$)



(a) Utility

(b) Cost

(c) Burnouts

(d) Responsibility Score

Fig. 6: Results for mixed population of ethical agents

have high utility and the lowest number of burnouts, while demonstrating highly responsible behaviour. These trends might change if the proportion of different types of ethical agents in the network is varied.

5 Discussion

We modelled ethical agents following a variety of paradigms of ethics, namely– utilitarianism, deontology, virtue ethics and transcendence. Utilitarian agents maximise 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. 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 of each other. We observe some commonalities across different models of ethics, which are discussed as follows. The highest transcendence level looks similar to utilitarianism, as agents at the 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 the case of transcendence, responsible behaviour is an emergent characteristic rather than something that agents are forced to uphold. Transcendence gives flexibility to agents to adapt to other agents and the environment based on changing contexts.

There are multiple models of ethics which can be modelled in autonomous agents. Each of these models of ethics has its nuance and reasoning to demonstrate ethical behaviour. SPECTRA as a framework provides a way to test, analyze and evaluate multiple models of ethics modelled in autonomous agents specifically in multi-agent settings. This enables the comparison of diverse models of ethics on a single testbed and also ensures that the system designer is aware of the details and implications of using different models of ethics. Thus, the SPECTRA framework can help a system designer evaluate different models of ethics specific to the scenario and context being modelled.

6 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 testbed 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 testbed can be extended in a variety of ways to incorporate different variations of autonomous agents and environments. 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.

Declarations - Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

[1] David Abel, James MacGlashan, and Michael L Littman. Reinforcement learning as a framework for ethical decision making. In *Workshops at the thirtieth AAAI conference on artificial intelligence*, 2016.

[2] Colin Allen, Iva Smit, and Wendell Wallach. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 7(3):149–155, 2005.

[3] Michael Anderson and Susan Leigh Anderson. Ethel: Toward a principled ethical eldercare robot. 2008.

[4] Michael Anderson and Susan Leigh Anderson. Geneth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics*, 9(1):337–357, 2018.

[5] Michael Anderson, Susan Leigh Anderson, and Chris Armen. Towards machine ethics. In *AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA*, 2004.

[6] Michael Anderson, Susan Leigh Anderson, and Chris Armen. Medethex: a prototype medical ethics advisor. In *AAAI*, pages 1759–1765, 2006.

[7] Nicolas Berberich and Klaus Diepold. The virtuous machine-old ethics for new technology? *arXiv preprint arXiv:1806.10322*, 2018.

[8] Selmer Bringsjord and Joshua Taylor. Introducing divine-command robot ethics. *Robot ethics: the ethical and social implication of robotics*, pages 85–108, 2012.

[9] Christopher Cloos. The utilibot project: An autonomous mobile robot based on utilitarianism. In *2005 AAAI Fall Symposium on Machine Ethics*, pages 38–45, 2005.

[10] Kari Gwen Coleman. Android arete: Toward a virtue ethic for computational agents. *Ethics and Information Technology*, 3(4):247–265, 2001.

[11] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14, 2016.

[12] Louise Abigail Dennis, Michael Fisher, and Alan Winfield. Towards verifiably ethical robot behaviour. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[13] Jayati Deshmukh, Nikitha Adivi, and Srinath Srinivasa. Modeling application scenarios for responsible autonomy using computational transcendence. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 2496–2498, 2023.

[14] Jayati Deshmukh, Nikitha Adivi, and Srinath Srinivasa. Resolving the dilemma of responsibility in multi-agent flow networks. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 76–87. Springer, 2023.

[15] Jayati Deshmukh and Srinath Srinivasa. Computational transcendence: Responsibility and agency. *Frontiers in Robotics and AI*, 9, 2022.

[16] Virginia Dignum. Responsible autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4698–4704, 2017.

[17] Virginia Dignum. Responsibility and artificial intelligence. *The Oxford Handbook of Ethics of AI*, 4698:215, 2020.

[18] Naveen Sundar Govindarajulu, Selmer Bringsjord, Rikhiya Ghosh, and Vasanth Sarathy. Toward the engineering of virtuous machines. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 29–35, 2019.

[19] Marcello Guarini. Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, 21(4):22–28, 2006.

[20] Marcello Guarini. Case classification, similarities, spaces of reasons, and coherences. In *Coherence: Insights from Philosophy, Jurisprudence and Artificial Intelligence*, pages 187–201. Springer, 2013.

[21] Marcello Guarini. Moral case classification and the nonlocality of reasons. *Topoi*, 32(2):267–289, 2013.

[22] Don Howard and Ioan Muntean. Artificial moral cognition: moral functionalism and autonomous moral agency. In *Philosophy and computing*, pages 121–159. Springer, 2017.

[23] Immanuel Kant and Jerome B Schneewind. *Groundwork for the Metaphysics of Morals*. Yale University Press, 2002.

[24] Seng W Loke. Designed to cooperate: a kant-inspired ethic of machine-to-machine cooperation. *AI and Ethics*, 3(3):991–996, 2023.

[25] Bertram F Malle, Matthias Scheutz, and Joseph L Austerweil. Networks of social and moral norms in human and robot agents. In *A world with robots*, pages 3–17. Springer, 2017.

[26] B McLaren. Lessons in machine ethics from the perspective of two computational models of ethical reasoning. In *2005 AAAI Fall Symposium on Machine Ethics*, pages 1–8, 2005.

[27] Bruce M McLaren. Extensionally defining principles and cases in ethics: An ai model. *Artificial Intelligence*, 150(1-2):145–181, 2003.

[28] Bruce M McLaren. Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE intelligent systems*, 21(4):29–37, 2006.

[29] Bruno Mermet and Gaële Simon. Formal verication of ethical properties in multiagent systems. In *1st Workshop on Ethics in the Design of Intelligent Agents*, 2016.

[30] Stefanie Meyer, Sarah Mandl, Dagmar Gesmann-Nuissl, and Anja Strobel. Responsibility in hybrid societies: concepts and terms. *AI and Ethics*, 3(1):25–48, 2023.

[31] John Stuart Mill. Utilitarianism. In *Seven masterpieces of philosophy*, pages 329–375. Routledge, 2016.

[32] John Stuart Mill and Jeremy Bentham. *Utilitarianism and other essays*. Penguin UK, 1987.

[33] BF dos S Neto, Viviane Torres da Silva, and Carlos JP de Lucena. Nbdi: An architecture for goal oriented normative agents. In *ICAART 2011*, 2011.

[34] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[35] Michal Pěchouček and Vladimír Mařík. Industrial deployment of multi-agent technologies: review and selected case studies. *Autonomous agents and multi-agent systems*, 17(3):397–431, 2008.

[36] Michal Pechoucek, Simon G Thompson, and Holger Voos. *Defence Industry Applications of Autonomous Agents and Multi-Agent Systems*. Springer, 2008.

[37] Gregory S Reed, Mikel D Petty, Nicholaos J Jones, Anthony W Morris, John P

Ballenger, and Harry S Delugach. A principles-based model of ethical considerations in military decision making. *The Journal of Defense Modeling and Simulation*, 13(2):195–211, 2016.

[38] Jaeeun Shim, Ronald Arkin, and Michael Pettinatti. An intervening ethical governor for a robot mediator in patient-caregiver relationship: Implementation and evaluation. In *2017 IEEE International conference on robotics and automation (ICRA)*, pages 2936–2942. IEEE, 2017.

[39] Torty Sivill. Ethical and statistical considerations in models of moral judgments. *Frontiers in Robotics and AI*, 6:39, 2019.

[40] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, 53(6):1–38, 2020.

[41] Chien Van Dang, Tin Trung Tran, Ki-Jong Gil, Yong-Bin Shin, Jae-Won Choi, Geon-Soo Park, and Jong-Wook 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)*, pages 155–158. IEEE, 2017.

[42] Ibo Van de Poel. The relation between forward-looking and backward-looking responsibility. In *Moral responsibility: Beyond free will and determinism*, pages 37–52. Springer, 2011.

[43] Dieter Vanderelst and Alan Winfield. An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 48:56–66, 2018.

[44] Ajay Vishwanath, Einar Duenger Bøhn, Ole-Christoffer Granmo, Charl Maree, and Christian Omlin. Towards artificial virtuous agents: games, dilemmas and machine learning. *AI and Ethics*, pages 1–10, 2022.

[45] 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, 2009.

[46] Alan FT Winfield, Christian Blum, and Wenguo Liu. Towards an ethical robot: internal models, consequences and ethical action selection. In *Conference towards autonomous robotic systems*, pages 85–96. Springer, 2014.

[47] Alan FT Winfield and Marina 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*, 376(2133):20180085, 2018.