The Triad of Identity, Trust and Responsibility in Multi-Agent Systems

Jayati Deshmukh
University of Southampton
Southampton, United Kingdom
j.deshmukh@soton.ac.uk

Vahid Yazdanpanah
University of Southampton
Southampton, United Kingdom
V.Yazdanpanah@soton.ac.uk

Sebastian Stein
University of Southampton
Southampton, United Kingdom
S.Stein@soton.ac.uk

Sarvapali D. Ramchurn
University of Southampton
Southampton, United Kingdom
sdr1@soton.ac.uk

ABSTRACT

The design of autonomous AI agents that behave responsibly and foster trust in open multi-agent systems remains a fundamental challenge. Traditional game-theoretical approaches largely assume self-interested behaviour, yet real-world collaborations among humans often rely on prosocial considerations that extend beyond individual utility. To address this, for the first time in this paper, we investigate the triad of *identity*, *responsibility*, and *trust* as core elements shaping responsible multi-agent behaviour. We propose a novel agent model, building on the notion of *Computational Transcendence*, which equips agents with an elastic sense of identity, enabling them to incorporate the welfare of others into their decision-making. Our framework integrates subjective (identity-based) and objective (experience-based and reputation-based) components of trust. Using Iterated Prisoner's Dilemma (IPD) simulations on different network structures, we analyse how varying levels of identity and trust affect responsible behaviour. Results demonstrate that the interplay of these three concepts can promote emergent responsibility, mitigate exploitation, and sustain long-term cooperation in dynamic multi-agent environments. We argue that this triadic perspective provides a principled foundation for designing trustworthy, responsible, and identity/value aware agents with implications for future human–AI collaboration.

KEYWORDS

Identity; Coordination; Responsibility; Trust; Ethics; Norms

ACM Reference Format:
Jayati Deshmukh, Vahid Yazdanpanah, Sebastian Stein, and Sarvapali D. Ramchurn. 2026. The Triad of Identity, Trust and Responsibility in Multi-Agent Systems. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/VTZX9616

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/VTZX9616

1 INTRODUCTION

For AI systems to collaborate with humans and ensure trustworthy collaborations, game-theoretical analysis provides a robust mathematical framework for modelling interactions. While traditional game theory approaches focus on modelling self-interested interacting agents, starting with the pioneering work of Axelrod [5] and Schelling [35], one open challenge has been to enrich game-theoretical models so that agents' interactions better reflect real-life practices of prosocial collaboration. In principle, as highlighted by [12], prosocial agents collaborate not purely for individual gain or utility, but also due to their subjective tendencies (what we call identity), which in game-theoretical terms could be modelled by incorporating the utility of others into their decision-making.

In real-life multi-agent systems that involve multiple decision makers, artificial (and human) agents interact with other agents to handle complex tasks. For example, robots coordinating with each other, language agents interacting to complete a shared task, and autonomous vehicles sharing information with others. Many of these settings are open-world systems where agents can join or leave the system at any point in time. Moreover, these agents may have different strategies based on which they can decide who, when and how to interact. In such settings, it becomes pertinent that the agents can trust other agents and act responsibly while interacting with others. For example, trust flows among a network of agents in settings like a supply chain, individual agents adapt their behaviour based on their interactions with agents, as in cyber security setting and multilateral interactions in realistic scenarios like climate change agreements. We discuss this in more detail towards the end of the paper.

The designers of such multi-agent systems want the system to reach a state (for example, robots or language agents to complete all the assigned tasks or autonomous vehicles to safely reach their destinations). There are broadly three ways to ensure that agents operating in such systems achieve the goal: a) a top-down approach, where the mechanisms and policies are designed to ensure that agents act in specific desirable ways aligned with the systemic interest. b) a bottom-up approach, when the agents learn by themselves and, based on their interactions with others, learn how to operate in a system. c) a hybrid approach, which combines both top-down and bottom-up approaches such that some aspects can be specified while some other aspects can be part of the agent model.

In these multi-agent systems, the agents might not know all the other agents they interact with and thus need a framework to trust them. Specifically, *trust* refers to the ability of an agent to handle the risk by delegating a task to another agent and expecting that it will make choices in the agent's interest. Trust has been broadly defined in two ways in multi-agent systems [32]: i) individual-level trust, which models different ways in which an individual agent trusts others based on its interactions with others. ii) system-level trust, which is used to ensure that agents act in a trustworthy manner. Also, there has been some work on modelling how trust propagates among different agents in a network [24, 25]. In this paper, we focus on modelling individual-level trust in agents using learning, reputation and cognitive-based approaches.

Responsible behaviour in multi-agent systems can be characterised as follows [50]: forward-looking responsibility, which enables agents to take actions to reach a desirable state and backward-looking responsibility, which ensures that agents take actions to avoid or prevent certain outcomes. Different techniques have been developed to enable agents to act responsibly, like machine ethics, artificial moral agents and normative agents [3, 4, 46]. In this paper, we build on an innate model of responsibility called Computational Transcendence (CT), which makes intrinsically responsible agents [12]. Specifically, in this work, we operationalise responsibility as the extent to which an agent incorporates the welfare of other agents into its decision-making, resulting in cooperative behaviour that improves collective outcomes over repeated interactions.

Unlike prior work on Computational Transcendence, which considered identity among directly connected agents, our model introduces a mechanism for trust propagation across indirect connections, enabling agents to incorporate trust-mediated identity beyond their immediate neighbours. The key contributions of this paper are as follows:

(a) We present an agent model which integrates identity and trust and leads to responsible behaviour by agents.
(b) We evaluate the proposed model in diverse network settings and parameters related to their representation of identity and trust to understand its impact on responsible behaviour.
(c) We provide a comparative analysis of our proposed model with baseline agent modelling approaches.
(d) Finally, we highlight the relevance of our model across diverse applications like supply chain, cybersecurity and international climate agreements.

This paper is organised in the following way. In Section 2, we elaborate on the background and present some of the key work that has been done in designing computational models of identity, responsibility and trust. We show how these three concepts connect conceptually to each other in Section 3. Next, in Section 4, we present our formulation of an agent having an elastic identity along with a way to model trust towards others, and we show that it results in responsible behaviour when interacting with others in a network. We present the simulation setup and experimental results in Section 5. Next, we discuss and highlight the relevance of our proposed approach in some real-world applications in Section 6. Finally, we conclude in Section 7 with the key takeaways and potential future directions of this work.

2 RESPONSIBILITY, TRUST AND IDENTITY

In this section, we will discuss some of the key work which has been done regarding computationally building models of responsibility, trust and identity in autonomous agents operating in multi-agent settings.

2.1 Models of Responsibility

There is a shared agreement on the need and importance of *responsible* AI [14]. However, responsibility is defined in diverse ways depending on the exact setting, like acting ethically, value alignment, social welfare, trustworthiness and accountability [43, 45].

Computational techniques like Artificial Moral Agents (AMA), Reflective Equilibrium (RE), and Value Sensitive Design (VSD) have been proposed to formally represent responsible behaviour in autonomous agents [19, 33, 46]. For the ethical or moral interpretation of responsible AI, there has been significant work around modelling normative ethics in agents so that they can act as per the appropriate paradigms of ethics, like utilitarianism, deontology or virtue ethics [1, 4, 11, 13, 41, 48].

In this context, the question of what it means for agents in a multi-agent system to act responsibly toward other human or AI decision makers remains largely unexplored. This issue is closely connected to the challenge of determining how, and to what extent, an AI agent should trust the decision makers with whom it interacts. Addressing this challenge (which is our focus in this work) not only requires an examination of the conditions under which trust is warranted, but also opens up the broader question of how trust can be formally related to, and integrated into, the modelling and analysis of responsible AI behaviour.

2.2 Models of Trust

Trust is an important characteristic of autonomous agents, based on which they can interact with each other in realistic ways [10, 17]. Trust allows agents to rely on other agents in the system, especially in the face of uncertainty. It also enables them to take calculated risks and delegate tasks to each other based on the underlying trust levels. It has been an active research area, and it has been explored and modelled in diverse ways. Also, it is an interdisciplinary topic and has been of significance in diverse fields like sociology, psychology, economics, computer science and artificial intelligence.

Social trust is beyond just trusting an agent to do something. It involves cognitive components like *cognitive evaluations* of belief about others; *belief about the context* which includes norms, institutional guarantees, social roles, sanctions and reputation mechanisms; expectations which involve understanding the consequences of trust and *goals* to accomplish something by trusting others [10, 17]. Another line of work has been along modelling trust in agents based on their reputation [21, 32]. The reputation of an agent can be estimated based on its past actions and alignment to its intentions and norms [30]. Settings with multiple agents interacting with each other involve aggregating the opinions of multiple agents. It has been used in diverse applications like wireless sensor networks, peer-to-peer networks, auctions, cloud providers, buyers and sellers on e-commerce platforms and access control.

This led to the formalisation of trust dynamics in multi-agent systems. It resulted in logic-based formal models which allow individual and groups of agents to reason about other agents in the system [15]. Specific circumstances or settings can impact an agent's ability to perform a task, despite having a good reputation in the past. This led to the development of *actual trust*, which is defined as the agents' capacity to deliver tasks in a specific context [2].

These models investigate and formalise trust in MAS, focusing on a relational view of how agents can reason about trust. While the impact of trust reasoning on decision making has been studied, its influence on whether AI systems exhibit responsible behaviour still requires both conceptual and experimental analysis. In our view, trust is a dual-natured concept. It depends not only on how other (potentially trustworthy) agents have treated the agent in the past, but also on the subjective identity and perception of the trusting agent, including how flexible they are in deciding whom to trust and to what extent [8].

2.3 Models of Identity

Humans have an intrinsic notion of an elastic sense of self or elastic identity, using which they identify with others. For example, a parent identifies with their children, a soldier identifies with their country, a team member identifies with the team and organisation. Broadly, approaches to computational modelling of agency or identity can be categorised as follows [40]: *normative*, *adaptive*, *quantitative*, and *autonomic* models of agency.

Normative models of agency model agents having a set of imperatives as well as discretionary entitlements. They also implement logical frameworks that encode different forms of individual and collective goals [9, 44, 49]. Adaptive frameworks can either learn the model based on its interactions or purely using positive or negative reinforcement signals from the environment [28, 37]. Next, agents can be modelled as per the rational choice theory [18, 31, 36]. In this approach, agents have a self-interest function, and they obtain a payoff resulting in a corresponding *utility*. Rational agents use utility maximisation to make decisions in this way. Finally, autonomic computing [20, 27] aims to provide computational entities with self-management properties, called "self-*" properties– like self-healing, self-tuning and self-recovery. The field of Autopoiesis was started by Maturana and Varela [29], who developed computational models of self-referential entities based on biological models of cognition.

There has been some work in the literature focusing on modelling identity in autonomous agents. However, there is a dearth of formal computational models which can formulate an abstract identity of autonomous agents such that they can operate in diverse settings. In this work, we build on the notion of Computational Transcendence [12], which endows an elastic identity to agents, using which they can computationally identify with other aspects in the system.

In this paper, we connect these three separate lines of work by proposing a computational model which combines identity, trust and responsibility in autonomous agents. We present the intuition of this approach and the formal model in the sections ahead.

3 FROM TRUST TO RESPONSIBILITY

In this paper, we focus on the triad of identity, responsibility and trust in multi-agent systems. Specifically, we focus on quantifying the extent of trust among the agents when they interact with each other repeatedly over time.

Individual level trust is defined as the extent to which an agent can rely on another agent, and based on this, it decides its interaction strategy with this agent [32]. Individual level trust has been modelled in three ways: i) *learning based*, where agents infer how others have interacted with it in the past and then it learns their behaviour over time. ii) *reputation-based*, where the agents share how other agents have interacted with them, and based on this collective opinion, agents form a notion of trust towards others. iii) *socio-cognitive based*, in which case agents try to infer the motivations of others based on their behaviour and then reasoning on those beliefs, estimate their trustworthiness score for the agents. On the other hand, in the field of psychology, trust has been defined as a subjective notion defined as the probability by which an agent expects another agent to act towards its welfare [17].

In this work, we combine both these approaches such that agents have a two aspects of trust: i) *a subjective element*, which characterises generally the extent to which agents identify and account for other agents in the system and, ii) *an objective element*, which uses learning and reputation based approaches to quantify trust based on the past interactions with itself and with others. We use a recently proposed approach called Computational Transcendence (CT) [12], which endows agents with an elastic sense of self, using which they can identify with others in the system. It has been shown that it results in emergent responsible behaviour by the agents. In this paper, we extend the CT model to incorporate trust and then evaluate the actions and behaviour of these agents.

Responsible behaviour in human–AI interactions is a complex and inherently normative concept [32, 38]. In principle, we expect individual agents (whether human or artificial) not only to pursue their own utility but also to consider, at least to some extent, the welfare of others. While such behaviour may appear counter-intuitive if we assume purely self-interested agents in single-shot interactions, long-term iterative interactions in agent societies reveal a different picture [7, 26]. In these settings, the utility of an individual depends on what can be achieved collectively, which motivates the need for models that can explain why collaboration emerges and for explanatory techniques capable of capturing such relations realistically. Modelling a notion of elastic identity in agents can be a potential solution for this [12]. Also, we would like to clarify that acting responsibly, factoring in others, is different from acting altruistically. In the first case, the agent maximises its utility while accounting for the other agents, while in the latter case, the agent foregoes itself and its utility, and acts for the benefit of the other. In this paper, we explore the former case where agents continue to act rationally and act for the collective welfare. This line of research was pioneered by Schelling [35], Axelrod's foundational work on collaboration [5] and philosophical foundations on norm emergence and normative coordination [22, 39, 42].

In this line of work, one approach is to view norms as intermediary mechanisms that help resolve the tension between individual and collective interests, e.g., see [23, 42]. While an agent considers

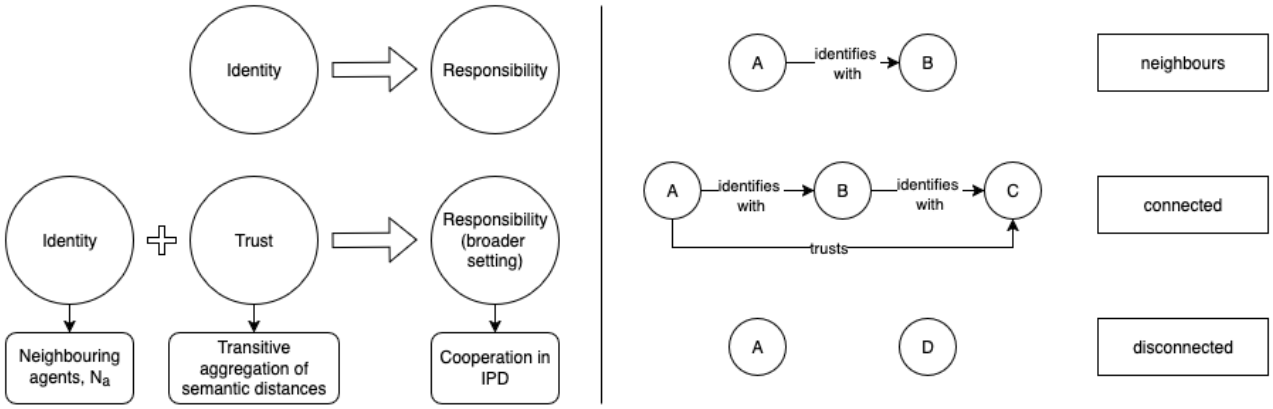


Figure 1: Interrelation between identity, trust and responsibility in multi-agent systems

its individual gain, this gain is not determined solely in isolation but also in relation to the collective benefit. Then, social norms, such as "be considerate of others", function as enabling tools that align individual decisions with collective welfare. This connects directly to Margalit's concept of Prisoner's Dilemma (PD) norms, which facilitate the resolution of social dilemmas and in an extended form for coordination (and coordination game scenarios[1]).

Building on this idea, one may conclude that agents should always cooperate as a whole (i.e., form the grand coalition in the game-theoretical sense). However, in real societies, individuals rely on both experiential factors (past interactions) and subjective or personal dispositions (e.g., upbringing, biological tendencies) to decide who they trust and whose utility matters to them. Thus, for an agent to operationalise a social norm of responsible behaviour, it must reason explicitly about trust: whom to trust, to what extent, and how to update trust over time. This motivates our central perspective in this work that:

> *Intelligent agents in a social multi-agent setting display responsible behaviour not only by pursuing individual gains but also by following social norms that take into account the welfare of those agents they trust, with trust itself evolving over time as the system dynamics unfold.*

To that end, this work puts forward the perspective that ensuring responsible behaviour of AI agents, particularly in human–AI collaborations, requires explicit modelling of trust: its scope, extent, and dynamics across different connections. We argue that this necessitates representing the social network of interactions, which allows one to distinguish between (1) agents an individual is merely able to interact with and (2) those it actually trusts, to varying degrees. Importantly, while the utility of others is an essential factor, an agent's own goals and identity (for example, whether it is generally open or closed towards others) remain central. Together, these considerations define the triadic interaction between identity, trust, and responsibility, which forms the basis of the dynamics we analyse in this work.

[1]Although in the remainder of this work we look into Prisoner's Dilemma settings with its standard characteristics, the idea to capture different aspects of trust and how this results in a more realistic explanations for responsible behaviour apply to other settings such as coordination games too.

Figure 1 shows the interrelation between identity, trust and responsibility, which is explored in this work. Some of the earlier work has shown that modelling an elastic identity in agents leads to responsible behaviour [12]. In this paper, we explore broader settings, where agents interact with diverse agents whom they trust to different extents. They identify with their neighbours but interact with everyone; they formulate a notion of trust towards other connected agents (whom they don't identify with) using our semantic distance aggregation techniques; and finally, we evaluate their level of responsibility based on the extent to which they cooperate with other agents in the Iterated Prisoner's Dilemma (IPD) setting. Specifically, we model three kinds of connections between agents: direct connections, indirect connections via multiple agents and disconnected agents. To investigate this hypothesis formally, we next introduce a computational model that integrates elastic identity with trust propagation in a networked multi-agent setting.

4 MODELLING GAME-THEORETICAL INTERACTIONS ON A SOCIAL GRAPH

In realistic settings, agents interact with different types of agents; they identify and care about some, indirectly know others based on their network, and are completely disconnected from others. In this paper, we extend the previously proposed model of Computational Transcendence [12], which demonstrated how agents act responsibly with other agents they identify with; we model a notion of trust linked to the identity among agents, which enables them to interact with agents with whom they are indirectly connected (but don't directly identify with). This is relevant and important, because in realistic settings it is more likely for agents to come across different types of connections: agents whom they identify with i.e. care about them and assist them in their goals; as well as agents whom they don't identify with (due to diverse reasons like these are newly added agents, they have had infrequent interactions with agents or just that the agents far away from them in the network).

It has been shown that the notion of trust is transitive and can flow through agents in a network [34]. Intuitively, if agent A trusts B, and B trusts C, then A can trust C not directly, but indirectly via B. Thus, in our proposed model, we combine the notions of identity

along with transitive trust and explore the extent to which it leads to responsible and cooperative behaviour by agents.

Our proposed approach has three main aspects: formulation of the agent, a, (with an identity and trust potentially leading to responsible behaviour), the utility function of the agent, $u(a)$, which specifies its utility depending on its choices as well as the choices of other agents in the multi-agent system, and finally, the network structure, G, of which the agent is part. Next, we elaborate on each of these in more detail. We model each agent a in the network with two parameters: first, a subjective element called *elasticity* denoted by γ_a, which represents the extent to which it identifies with other agents in the network, specifically its direct 1-hop neighbours; and second, an objective element called *semantic distance* denoted by $d_a(b)$ which quantifies the relative semantic distance of agent a towards one of its neighbouring agent b in the network. Elasticity, γ_a, is an agent-level feature which signifies the extent to which an agent factors others' utility while making decisions. A low elasticity value denotes a selfish agent which predominantly cares about its own utility, and a high elasticity value denotes an agent which highly accounts for others. On the other hand, semantic distance, $d_a(b)$, is an edge-level feature, which denotes the proportion with which the agent identifies with other agents. Its low value denotes higher identification, while a higher value denotes lower identification. As discussed in detail ahead, the agent scales the utility by a factor of γ^d when accounting for the utility of other agents it identifies with.

Such an agent derives utility not only from its payoffs but also from the payoffs of other agents that it identifies with. Generally, it applies to all aspects that the agent identifies with; however, in this paper, we represent the identity set of an agent by its neighbouring agents. Different variables of interest are denoted as follows: agent a has an identity set I_a which consists of all its 1-hop neighbours N_a, semantic distance $d_a(b) \rightarrow R^+$ denotes the directed distance of agent a towards its neighbouring agent b. Agent a derives a scaled utility based on its interactions with neighbouring agent b, which gets a payoff $\pi_i(b)$ in game state i. In such a setting, the utility of agent a is computed as follows:

$$u_i(a) = \frac{1}{1 + \sum_{\forall b \in N_a} \gamma_a^{d_a(b)}} \left(\pi_i(a) + \sum_{\forall b \in N_a} \gamma_a^{d_a(b)} \cdot \pi_i(b) \right) \quad (1)$$

This formulation ensures that an agent's utility reflects a weighted combination of its own payoff and those of agents it identifies with (and the weights depend on both its elasticity and semantic distance). After every few epochs, agents update semantic distances to their neighbours based on their interactions with them. Agents calculate the proportion of rewards, $r_a(b)$ and costs, $c_a(b)$, from each neighbour relative to the aggregated rewards and costs, and then update the distance by the difference between the proportional reward and the cost. Reward, $r_a(b)$ represents the part of the utility, $u(a)$ derived from the interaction with b. In the current simulation, the costs are negligible and thus can be ignored. The semantic distance update equation is defined as follows:

$$\begin{aligned} \Delta d_a(b) &= \frac{e^{r_a(b)}}{\sum_{\forall b \in N_a} e^{|r_a(b)|}} - \frac{e^{c_a(b)}}{\sum_{\forall b \in N_a} e^{|c_a(b)|}} \\ d_a(b) &= d_a(b) - \lambda \Delta d_a(b) \end{aligned} \quad (2)$$

In prior work, interactions were limited to neighbours, i.e. agents interacted with only those agents they identified with. In this paper, we extend the CT model to broader, more realistic settings where agents interact with neighbouring agents, connected but non-neighbouring agents and disconnected agents. Agents have a semantic distance as discussed above to each of their direct neighbours. However, for connected non-neighbouring agents and disconnected agents, there is no distance parameter since agents do not directly identify with these agents. We posit that by incorporating a notion of trust in these agents, we can also enable them to interact with agents they do not directly identify with responsibly. The agents estimate the distance of non-neighbouring connected agents by aggregating all the pair-wise distances denoted by x_i between the two agents as follows:

$$d_a^+(b) = \left(\prod_{i=a}^{b} x_i \right)^{\frac{1}{\alpha \cdot n}} \quad (3)$$

Here, we take the product of all the individual distances on the path (path with the least number of hops) between agent a and b. To address the impact of variable path length, it is powered to $1/\alpha \cdot n$, where n is the shortest number of hops between the agents and α is a scaling constant. To simplify the computation, it can be rewritten as follows:

$$d_a^+(b) = \exp\left(\frac{1}{\alpha \cdot n} \sum_{i=a}^{b} \log x_i \right) \quad (4)$$

Thus a transcended agent a can have the distance $d_a(b)$ to other agent b as per the following three cases:

$$d_a(b) = \begin{cases} d_a(b) & \text{if a and b are neighbours,} \\ d_a^+(b) & \text{if a and b are connected but non-neighbours,} \\ \infty & \text{if a and b are disconnected.} \end{cases} \quad (5)$$

Finally, the overall formulation is a multi-agent system where the agents are placed in a network. Specifically, the setup can be represented as a network of agents $G = (V, E)$ modelled as an undirected graph having V nodes representing the agents and E representing the directly connected neighbours or the identity set of agents. Every agent interacts with every other agent in the system; neighbours, connected and disconnected agents, over multiple epochs.

5 SIMULATION RESULTS AND DISCUSSION

Prisoner's Dilemma is a well-known game in game theory which denotes the consequences of agents acting only based on their self-interest. As shown in Table 1, we have two agents, A and B, and both of them have two choices: either to cooperate, C or defect, D, with each other. If they both cooperate, then they both get a payoff

		Agent A	
		C	D
Agent B	C	R = 6, R = 6	S = 0, T = 10
	D	T = 10, S = 0	P = 1, P = 1

Table 1: Payoff Matrix for 2-player Prisoner's Dilemma

of R=6. However, as long as the other agent cooperates, there is an incentive for the agent to switch to defecting and gain a higher payoff of T=10 while the other agent still cooperating gets S=0. That makes the other agent also choose to defect, and the game ends in a DD state where both agents get a payoff of P=1. This state is the Nash Equilibrium of the game, and defection is the strictly dominant strategy of the game since, irrespective of what choice the other agent makes, it is always better for an agent to defect as compared to cooperate.

In this game, choosing to cooperate can be termed as the *responsible* choice since if both agents cooperate, it results in a better outcome for both of them. However, it is very difficult to ensure that the agents reach the CC state of the game due to the temptation to defect. On the other hand, defection can be termed as the *irresponsible* choice, since it results in either a large disparity in payoffs (in CD or DC state) or very low payoff for both the agents (in DD state). And yet, agents might choose to defect since it is the *rational* choice.

One of the ways to enable cooperation in this game is to make the agents play the prisoner's dilemma with each other repeatedly, over multiple iterations [5]. It is called the Iterated Prisoner's Dilemma (IPD). In a one-shot game, the agents are more likely to defect since there are no long-term consequences of their choice. However, in a repeated setting, there is a notion of *memory* by which agents remember the choices of other agents in the earlier rounds. Thus, the IPD setting brings in long-term consequences of the agents' choices as the game progresses over multiple epochs. And this in turn, leads to more cooperative and responsible behaviour by the agents.

In our simulation, agents play Iterated Prisoner's Dilemma (IPD) with all other agents over multiple epochs, and after every game, the agents receive payoffs based on the game state as shown in Table 1. We simulate the game for some epochs, then the agents update their semantic distances to their neighbours till it settles (when the changes in semantic distances are $< \epsilon = 0.01$), and then we run the simulation again for a fixed number of epochs and present those results.

The simulation is developed using the NetworkX library in Python[2]. We generate a random network (Erdős–Rényi network) of 25 agents. The agents play IPD with every other agent in the network for 100 rounds. After which, the agents update their semantic distances to each other based on their interactions. Next, we repeat the simulation for 100 rounds with the updated semantic distances. For each setting, the simulations are run for 10 different random seeds (in order to avoid outlier results due to a specific seed), and the aggregated results over multiple runs are presented. Next, some of the simulation results are presented in further detail.

[2]https://networkx.org/

5.1 Compare proposed approach with other baseline strategies

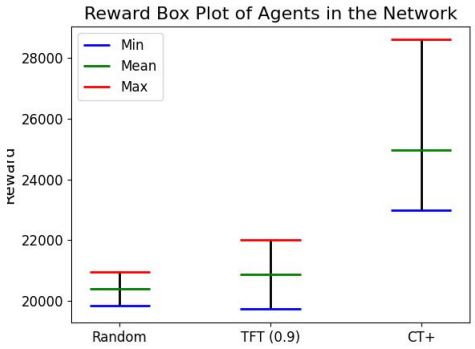


Figure 2: Reward distribution across different strategies

First, we start by comparing our proposed approach with other baseline agent strategies. We initialise all the agents with the selected strategy and let the simulation run for the maximum number of epochs. The network structure remains the same across all the simulation runs. The simplest strategy is the Random strategy, where all the agents in the network choose with a random probability of 0.5 whether to cooperate or defect with the other agent. In this case, each game is independent and does not depend on past or future games. Next, we simulate the popular Tit-for-Tat (TFT) strategy. All the agents start with cooperation and continue to do so when everyone has a TFT strategy. So we introduce a 10% randomness such that agents respond as per the TFT strategy 90% of the time and respond randomly for the rest 10%. This brings some variation in the game.

We compare our proposed approach called *CT+* with the two strategies Random and TFT (0.9) as shown in Figure 2. We plot the minimum, maximum and average reward of all the agents in the network. In terms of the value of rewards, we note that CT+ agents get a higher reward than the other two types of baseline agents.

Figure 3 shows the proportion of times the game is one of the three game states over all the runs across all the agents in the network. Any game can be in either CC, CD (includes both CD

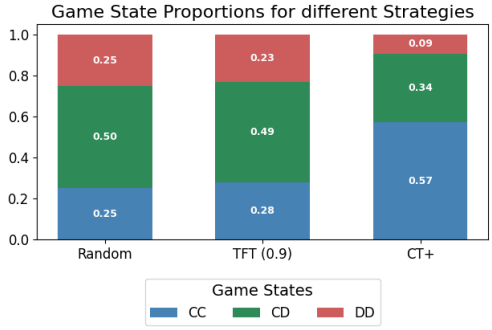
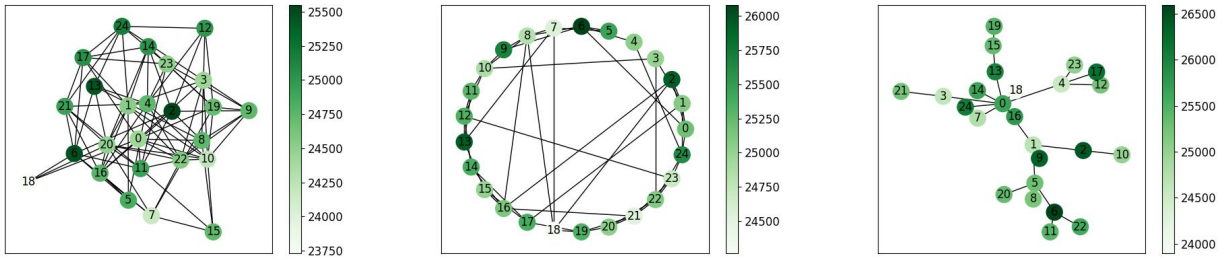
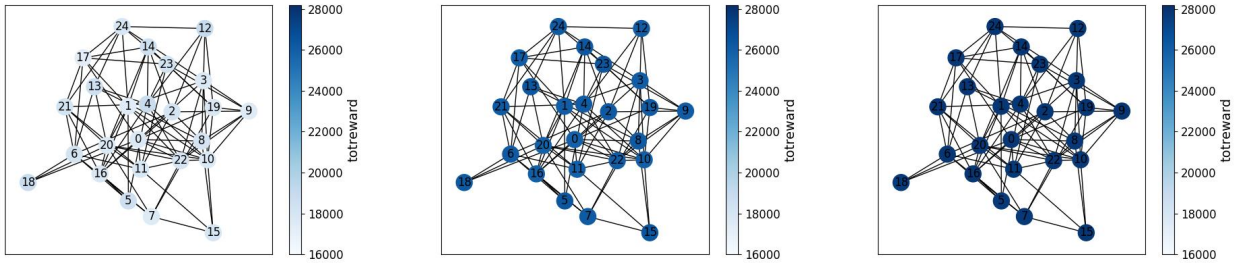


Figure 3: Distribution of states across different strategies



(a) Erdős–Rényi Network (NP:3.06e-18) (b) Watts–Strogatz Network (NP:2.98e-18) (c) Barabási–Albert Network (NP:2.64e-18)

Figure 4: Distribution of utility and its Nash Product (NP) for Erdős–Rényi, Watts–Strogatz and Barabási–Albert networks



(a) Low elasticity (b) Medium elasticity (c) High elasticity

Figure 5: Results for low, medium and high elasticity (γ)

and DC) or DD game state. We observe that Random and TFT (0.9) have similar distributions across the three game states. While the distribution of game states for CT+ is different, where almost 60% of the time the game ends in the CC state. These results demonstrate that integrating identity and trust enables agents to achieve higher collective welfare and more equitable outcomes compared to baseline strategies.

5.2 Simulations in different types of networks

Next, we explore the impact of different networks between the agents. We hypothesise that network structure affects the interaction with connected agents since the distance is estimated based on the individual distances of the shortest path.

We explore the following three types of networks: a) Erdős–Rényi network [16]: It is a random network, which first generates all the nodes and then every pair of nodes is either connected or disconnected with a probability p. The probability of each edge is independent of the other edges in the network. b) Watts–Strogatz network [47]: It is a network which has small-world properties like small path lengths and high clustering coefficients, which are observed in many real-world networks. It models triadic closures, i.e. if an agent is connected to two different agents, then there is a high probability that those two agents are also connected to each other. In order to build this network, first, all the agents are arranged in a circle with k connections on each side, and next, some of these connections are removed and connected to a random agent in the network. c) Barabási–Albert network [6]: It is a network which has a scale-free property and is generated using preferential attachment. This network is generated by adding agents one by one to the network. Every time a new agent is added, it connects to one of the existing agents in the network proportional to the degree of that agent. It results in the creation of a few hubs having a high degree and a lot of spokes having a low degree.

Figure 4 shows the results of our proposed approach on the three types of networks. We note that the rewards are fairly distributed in the Erdős–Rényi network with the highest Nash Product (NP) of normalised agent utilities. In the Watts–Strogatz network and Barabási–Albert network, we observe that there are a few agents with high rewards and a lot of agents with very low rewards. In future, we will also explore other alternatives to selecting a path with the lowest number of hops, such as the most reliable path or the minimum distance path, which is relevant in networks with multiple paths between agents, like the Erdős–Rényi or Watts–Strogatz network. For the remaining simulations, we utilise the Erdős–Rényi network to model the network structure among the agents.

5.3 Simulation for varying elasticity

Next, we vary the elasticity of the agents and evaluate the extent to which it affects the choices of the agents. To recap, elasticity is the extent to which agents identify with other agents, where a

low value denotes predominantly selfish agents and a high value denotes agents which highly identify and care about others. We simulate agents with three elasticity levels, which is randomly sampled as follows: low elasticity in $[0.05, 0.35]$, medium elasticity in $[0.35, 0.65]$ and high elasticity in $[0.65, 0.95]$.

Figure 5 shows the variation in average agent rewards for low, medium and high elasticity settings. We keep the colour scales fixed across the three plots to ensure that the results are comparable. As the elasticity increases, agents act more cooperatively and responsibly, and this in turn, leads to higher rewards. The average reward per agent for low elasticity is 18031, for medium elasticity it is 26125 and for high elasticity it is 27803. Thus, an interesting observation here is that agents perform well and receive significantly high rewards even when operating with medium elasticity. Additionally, the proportion of games ending in the CC state increases from 22% to 64% to 83% for low, medium and high elasticity, respectively.

6 THE TRIAD IN PRACTICE

While in earlier sections, we showed how this perspective on identity and trust modelling results in collaborative responsible behaviour in experimental settings, this section discusses three scenarios with specific characteristics. We first look at a supply chain scenario and show how our approach allows for modelling "*indirect trust*", then look into a cyber-security example where the presented model shows "*adaptivity and tolerance*" (against malicious agents), and finally discuss a climate change agreement scenario to show how our approach expands (from abstract 1-1 prisoners' dilemma cases) to responsible behaviour in "*multilateral collaborations*".

6.1 From Direct to Indirect Trusting

Let's take a simple example of three suppliers in a supply chain, S_A, S_B and S_C. Assume that supplier S_A supplies a product to S_B, who further processes it and then dispatches it to S_C. In this scenario, since S_A, S_B and S_B, S_C directly interact with each other, they build a relation over time and thus they trust each other. In case of occasional exceptions like delayed payments or defects in products, they trust each other and respond accordingly.

In the same setting, let's say S_C needs some products from S_A. It has not directly interacted with S_A; however, S_B conveys whether S_A is a good supplier or not. Based on this information, S_C can make a more informed decision regarding how to interact with S_A.

Our proposed approach can be useful in this process. It helps agents to formally convey how their interactions have been with the agents they identify and directly interact with in terms of their semantic distances towards them at the current instant. And based on this information, we can estimate transitive trust, based on which agents can take better, well-informed decisions.

6.2 Adaptively Responsible Agents

Next, let's take the example of agents in cybersecurity. Let there be the following two types of agents: attacker agent, A_A, who tries to exploit network vulnerabilities and defender agent, A_D, who tries to learn to predict and block attacks. The interaction between these two agents is defensive and different from the interaction among the agents in the prisoner's dilemma.

In this scenario, the distance update feature of our approach is useful. Initially, the defender agent, A_D, can start by being sceptical about A_A and trust it by 0.5. Then, based on its interactions with A_A, it can update its semantic distance towards it; it can increase its semantic distance if A_A acts maliciously and reduce its semantic distance if A_A acts non-maliciously.

Also, since the semantic distances are updated after every few interactions, even if A_A is intentionally changing its behaviour from malicious to non-malicious or vice versa, the model can easily identify and respond to such tactics.

This example highlights the following key features of our proposed approach. First, adaptivity, as the model continuously adapts the semantic distances based on its interactions with the agents. Also, the need to label agents in the network as 'malicious' or 'non-malicious' is avoided. Instead, the model can infer the type of agent based on its neighbours' interactions with this agent over time. Second, resilience, since the agents cannot game the model, for example, by being non-malicious in the beginning to sway the model's opinion. Third, it works without any prior information or with partial information regarding the agents. And fourth, the network of defender agents can share the details regarding their interactions with other agents over time in the trusted network.

6.3 Scalability to Multilateral Collaborations

Finally, we demonstrate that the Iterated Prisoner's Dilemma (IPD), which has been used to illustrate our approach, is a setting which commonly occurs in multiple multi-lateral collaborations. We demonstrate this using the example of multiple countries taking steps towards climate change. Climate is a "global common", no one owns the climate, everyone benefits from a stable climate, but everyone can contribute to its degradation through greenhouse gas (GHG) emissions. Each country gains short-term economic benefits by emitting (for example, through industrial growth, energy access, and transportation). But the costs (for example, warming, extreme weather, biodiversity loss) are shared globally.

Our proposed approach can be used by countries to track their interactions individually, as well as evaluate the actions taken by countries impacting climate change over time. They can also share these details with other trusted countries. Identity and trust among the countries can be one way of enabling them to act responsibly towards the global climate change problem.

7 CONCLUSIONS

In this paper, we introduced an agent model that integrates identity and trust to promote responsible behaviour in multi-agent systems. Through evaluations across diverse network conditions, the model demonstrates its capacity to enhance cooperation and responsibility among agents. Furthermore, the model's applicability to domains such as supply chain management, cybersecurity, and international climate governance underscores its broader relevance. Currently, we focus on a specific understanding of "responsible behaviour", and this motivates looking into using this triad framework to explore its impact on other ethical principles. Overall, this work provides a principled foundation for advancing responsible and trustworthy agent design in complex socio-technical environments.

ACKNOWLEDGMENTS

This work is supported by Turing AI Fellowship (EP/V022067/1) on Citizen-Centric AI Systems (https://ccais.ac.uk/) and by Responsible Ai UK (EP/Y009800/1) (https://rai.ac.uk/).

REFERENCES

[1] Nirav Ajmeri, Hui Guo, Pradeep K Murukannaiah, and Munindar P Singh. 2020. Elessar: Ethics in norm-aware agents. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems*. 16–24.

[2] Michael Akintunde, Vahid Yazdanpanah, Asieh Salehi Fathabadi, Corina Cirstea, Mehdi Dastani, and Luc Moreau. 2024. Formal specification of actual trust in multiagent systems. In *3rd International Conference on Hybrid Human-Artificial Intelligence, HHAI 2024*. IOS Press, 22–35.

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

[4] Giulia Andrighetto, Guido Governatori, Pablo Noriega, and Leendert WN van der Torre. 2013. *Normative multi-agent systems*. Vol. 4. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[5] Robert Axelrod and William D Hamilton. 1981. The evolution of cooperation. *science* 211, 4489 (1981), 1390–1396.

[6] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[7] Rutger Bregman. 2020. *Humankind: A hopeful history*. Bloomsbury Publishing.

[8] Ningmeng Cao, Binghai Sun, Weijian Li, and Guoan Yue. 2025. The implicit theories of trust: the more individuals believe trust to be unchangeable, the more they tend to trust others. *BMC psychology* 13, 1 (2025), 950.

[9] Cristiano Castelfranchi, Frank Dignum, Catholijn M Jonker, and Jan Treur. 1999. Deliberative normative agents: Principles and architecture. In *International Workshop on Agent Theories, Architectures, and Languages*. Springer, 364–378.

[10] Cristiano Castelfranchi and Rino Falcone. 1998. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In *Proceedings International Conference on Multi Agent Systems (Cat. No. 98EX160)*. IEEE, 72–79.

[11] Amit K Chopra and Munindar P Singh. 2018. Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 48–53.

[12] Jayati Deshmukh and Srinath Srinivasa. 2022. Computational transcendence: Responsibility and agency. *Frontiers in Robotics and AI* 9 (2022), 977303.

[13] Jayati Deshmukh, Vahid Yazdanpanah, Sebastian Stein, and Timothy J Norman. 2024. Ethical alignment in citizen-centric ai. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 43–55.

[14] Virginia Dignum. 2019. *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Vol. 2156. Springer.

[15] Nagat Drawel, Jamal Bentahar, Amine Laarej, and Gaith Rjoub. 2022. Formal verification of group and propagated trust in multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 19.

[16] Paul Erdős and Alfréd Rényi. 1959. On random graphs I. *Publicationes Mathematicae* 6, 290-297 (1959), 18.

[17] Rino Falcone and Cristiano Castelfranchi. 2001. Social trust: A cognitive approach. In *Trust and deception in virtual societies*. Springer, 55–90.

[18] Jacques Ferber and Gerhard Weiss. 1999. *Multi-agent systems: an introduction to distributed artificial intelligence*. Vol. 1. Addison-Wesley Reading.

[19] Batya Friedman, Peter Kahn, and Alan Borning. 2002. Value sensitive design: Theory and methods. *University of Washington technical report* 2-12 (2002).

[20] Alan G Ganek and Thomas A Corbi. 2003. The dawning of the autonomic computing era. *IBM systems Journal* 42, 1 (2003), 5–18.

[21] Jones Granatyr, Vanderson Botelho, Otto Robert Lessing, Edson Emílio Scalabrin, Jean-Paul Barthès, and Fabrício Enembreck. 2015. Trust and reputation models for multiagent systems. *ACM Computing Surveys (CSUR)* 48, 2 (2015), 1–42.

[22] Davide Grossi and Frank Dignum. 2004. From abstract to concrete norms in agent institutions. In *International Workshop on Formal Approaches to Agent-Based Systems*. Springer, 12–29.

[23] Davide Grossi, Luca Tummolini, and Paolo Turrini. 2012. Norms in game theory. In *Agreement technologies*. Springer, 191–197.

[24] Ramanthan Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2004. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*. 403–412.

[25] Chung-Wei Hang, Yonghong Wang, and Munindar P Singh. 2009. Operators for propagating trust and their evaluation in social networks. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 1025–1032.

[26] Brian Hare and Vanessa Woods. 2021. *Survival of the friendliest: Understanding our origins and rediscovering our common humanity*. Random House Trade Paperbacks.

[27] Jeffrey O Kephart and David M Chess. 2003. The vision of autonomic computing. *Computer* 1 (2003), 41–50.

[28] Charles M Macal and Michael J North. 2005. Tutorial on agent-based modeling and simulation. In *Proceedings of the Winter Simulation Conference, 2005*. IEEE, 14–pp.

[29] Humberto R Maturana and Francisco J Varela. 1991. *Autopoiesis and cognition: The realization of the living*. Vol. 42. Springer Science & Business Media.

[30] Lik Mui, Mojdeh Mohtashemi, and Ari Halberstadt. 2002. A computational model of trust and reputation. In *Proceedings of the 35th annual Hawaii international conference on system sciences*. IEEE, 2431–2439.

[31] Simon Parsons and Michael Wooldridge. 2002. Game theory and decision theory in multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 5, 3 (2002), 243–254.

[32] Sarvapali D Ramchurn, Dong Huynh, and Nicholas R Jennings. 2004. Trust in multi-agent systems. *The knowledge engineering review* 19, 1 (2004), 1–25.

[33] John Rawls. 2009. *A theory of justice*. Harvard university press.

[34] Oliver Richters and Tiago P Peixoto. 2011. Trust transitivity in social networks. *PloS one* 6, 4 (2011), e18384.

[35] Thomas C Schelling. 1980. *The Strategy of Conflict: with a new Preface by the Author*. Harvard university press.

[36] Elham Semsar-Kazerooni and Khashayar Khorasani. 2009. Multi-agent team cooperation: A game theory approach. *Automatica* 45, 10 (2009), 2205–2213.

[37] Yoav Shoham, Rob Powers, and Trond Grenager. 2003. Multi-agent reinforcement learning: a critical survey. *Web manuscript* (2003).

[38] Munindar Singh. 1992. On the semantics of protocols among distributed intelligent agents. In *Eleventh Annual International Phoenix Conference on Computers and Communication [1992 Conference Proceedings]*. IEEE, 379–386.

[39] Munindar P Singh. 2014. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 1 (2014), 1–23.

[40] Srinath Srinivasa and Jayati Deshmukh. 2020. The Evolution of Computational Agency. In *Novel Approaches to Information Systems Design*. IGI Global, 1–19. https://doi.org/10.4018/978-1-7998-2975-1.ch001

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

[42] Edna Ullmann-Margalit. 2015. *The emergence of norms*. OUP Oxford.

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

[44] Wiebe Van der Hoek and Michael Wooldridge. 2003. Towards a logic of rational agency. *Logic Journal of IGPL* 11, 2 (2003), 135–159.

[45] Nicole A Vincent. 2011. A structured taxonomy of responsibility concepts. In *Moral responsibility: Beyond free will and determinism*. Springer, 15–35.

[46] Wendell Wallach and Colin Allen. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

[47] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *nature* 393, 6684 (1998), 440–442.

[48] Jessica Woodgate and Nirav Ajmeri. 2024. Macro ethics principles for responsible AI systems: Taxonomy and directions. *Comput. Surveys* 56, 11 (2024), 1–37.

[49] Fabiola López y López, Michael Luck, and Mark d'Inverno. 2006. A normative framework for agent-based systems. *Computational & Mathematical Organization Theory* 12, 2-3 (2006), 227–250.

[50] Vahid Yazdanpanah, Enrico H Gerding, Sebastian Stein, Corina Cirstea, MC Schraefel, Timothy J Norman, and Nicholas R Jennings. 2021. Different forms of responsibility in multiagent systems: Sociotechnical characteristics and requirements. *IEEE Internet Computing* 25, 6 (2021), 15–22.