JASSS
THE JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION

Using Agent-Based Modelling and Reinforcement Learning to Study Hybrid Threats

Kärt Padur[1], **Hervé Borrion**[2], **Stephen Hailes**[3]

[1] *Department of Computer Science, University College London, 66-72 Gower Street, WC1E 6EA, London, United Kingdom*
[2] *Department of Security and Crime Science, University College London, 35 Tavistock Square, WC1H 9EZ, London, United Kingdom*
[3] *Department of Computer Science, University College London, 66-72 Gower Street, WC1E 6EA, London, United Kingdom*
Correspondence should be addressed to *kart.padur.20@ucl.ac.uk*

Journal of Artificial Societies and Social Simulation 28(1) 1, 2025
Doi: 10.18564/jasss.5539 Url: http://jasss.soc.surrey.ac.uk/28/1/1.html

Received: 16-07-2024 Accepted: 06-11-2024 Published: 31-01-2025

Abstract: Hybrid attacks coordinate the exploitation of vulnerabilities across domains to undermine trust in authorities and cause social unrest. Whilst such attacks have primarily been seen in active conflict zones, there is growing concern about the potential harm that can be caused by hybrid attacks more generally and a desire to discover how better to identify and react to them. In addressing such threats, it is important to be able to identify and understand an adversary's behaviour. Game theory is the approach predominantly used in security and defence literature for this purpose. However, the underlying rationality assumption, the equilibrium concept of game theory, as well as the need to make simplifying assumptions can limit its use in the study of emerging threats. To study hybrid threats, we present a novel agent-based model in which, for the first time, agents use reinforcement learning to inform their decisions. This model allows us to investigate the behavioural strategies of threat agents with hybrid attack capabilities as well as their broader impact on the behaviours and opinions of other agents. In this paper, we demonstrate the face validity of this approach and argue that its generality and adaptability render it an important tool in formulating holistic responses to hybrid threats, including proactive vulnerability identification, which does not necessarily emerge by considering the multiple threat vectors independently.

Keywords: Hybrid Threats, Agent-Based Modelling, Reinforcement Learning, Cyberattack, Misinformation

Introduction

1.1 The topic of hybrid threats has gained significant attention in recent years, particularly after Russia attacked Ukraine in 2014 (Treverton et al. 2018; Sazonov et al. 2016). During this attack, Russia coordinated actions across multiple domains. For example, they deployed unmarked soldiers, engaged local separatists, conducted cyberattacks, and launched misinformation campaigns to achieve their strategic objectives. Russia did not trigger a significant international response to these attacks. Other historical examples include: state actors, such as Iran (Hoffman 2010) and China (Treverton et al. 2018); and non-state actors, such as Hezbollah (Azani 2013) and the Islamic State in Iraq and the Levant (Jacobs & Samaan 2015).

1.2 In recent years, hybrid operations have changed in scale by expanding beyond the physical space (Hoffman 2010) into the digital space due to the increasingly widespread use of the Internet and social media (Treverton 2018). Planning and launching such operations is typically less expensive than engaging in traditional warfare. Moreover, definitively attributing such attacks can be challenging, making a timely and appropriate response difficult, if not almost impossible, for the target (Treverton 2018). These attack methods offer many opportunities to manipulate people's attitudes and opinions and influence their decision-making (Linkov et al. 2019).

1.3 In this research, we define hybrid threats as *strategically coordinated actions that exploit vulnerabilities across cyber and information domains to influence the behaviour and shape the opinions of the target audience while staying below the threshold of effective response*.

1.4 To respond to the societal challenges associated with hybrid threats, it is essential to be able to identify and understand the attacker's behaviour. Modelling the behaviour of an adversary capable of coordinating attacks across multiple domains helps to understand its potential attack strategies and assess its likely impacts. Game theory is widely used in security and defence science research for this purpose. It has been used to model the behaviour of an attacker with specific attack capabilities, such as cyberattacks against cyber-physical systems (Huang & Zhu 2020) and misinformation dissemination in social networks (Kumar & Geethakumari 2013), as well as coordinated attack capabilities (Keith & Ahner 2021; Balcaen et al. 2022). However, the underlying rationality assumption (Mueller 1996) and equilibrium concept (Smith & Price 1973), as well as the need for a considerable degree of abstraction, can limit the use of game theory in the study of such emerging threats.

1.5 To address these limitations, we propose a novel agent-based model in which agents use reinforcement learning (RL) to make decisions. For this, we introduce agents into an environment that comprises a series of competing service providers and their customers. It also includes a social network in which information – both accurate and misleading – about the quality of service can be exchanged. In this environment, malicious agents use RL to decide when and for how long to attack a service provider, as well as how to coordinate attacks that exploit vulnerabilities across cyber and information domains. Other agents, with no malicious intent, use RL to decide from which provider to request service or about whom to exchange information. We use modelling and simulation to analyse the strategic behaviour of malicious agents with hybrid attack capabilities as well as their impact on the behaviours and opinions of other agents. The contributions of this paper are summarised as follows:

- We define a model with which to study the behaviour of malicious agents with hybrid attack capabilities. To the best of our knowledge, this is the first paper to study the behavioural strategies of malicious agents that exploit vulnerabilities across cyber and information domains.
- We represent malicious agents' behaviour as an RL problem to understand how adversaries can develop their attack strategy based on their experience with and observations of the dynamics of the surrounding system.
- We evaluate the impact of such attacks on the behaviours and opinions of the targets.

1.6 The problem is multifaceted and, whilst game theory can be applied to parts of it, the multiple layers of interaction render this approach inadequate to the task. Consequently, we propose that using RL in this domain is appropriate, effective, and easier to adapt to the complexities of this emerging threat.

1.7 The remainder of the paper is organised as follows: Section 2 introduces hybrid threats and relevant research on modelling the concept. Section 3 first introduces the environment before presenting different types of agents. The section then summarises the model. Section 4 details the design of the experiments and presents the results. Section 5 interprets the results and discusses the implications. Section 6 details the conclusions, limitations, and future work.

Related Work

2.1 The section introduces the concept of hybrid threats before describing the related research on modelling attacker behaviour.

Hybrid threats

2.2 There is no concrete definition of the term "hybrid threats" that has been agreed upon in policy documents and academic publications. Both state and non-state actors can use hybrid tactics to achieve their goals (Linkov et al. 2019). Rather than targeting armies, they typically target members of society to achieve their desired outcomes without escalating the situation into war (Gunneriusson & Ottis 2013; Treverton et al. 2018). They simultaneously use a variety of political, economic, social, and informational means to exploit vulnerabilities across different domains (Hoffman 2007; Treverton et al. 2018), making detection, attribution and response more challenging for the target (NATO Strategic Communications 2020).

2.3 Hybrid threats have been attracting growing interest among governments and policy-makers (UK Government 2015; Giannopoulos et al. 2020), inter-governmental organisations (European Commission 2016), defence and security communities (NATO Strategic Communications 2020), and academia (Hoffman 2010; Linkov et al. 2019). The reasons are the changing nature of conflict (Hoffman 2010), increasing interconnectedness between digital systems (Linkov et al. 2019), emerging technologies, and the changing information environment (Treverton et al. 2018).

Modelling hybrid threats

2.4 Agent-based modelling is a computational approach to modelling systems with many autonomous agents (Sayama 2015). Formal descriptions of interactions and their consequences, typically game theory and reinforcement learning, underpin the way in which computational agents interact with each other and seek to achieve their goals in such systems.

2.5 Game theory is a method that has been used to address different societal challenges, such as cooperation, conflict, and coordination (Myerson 1991; Poole & Mackworth 2017). It provides an accessible and widely-used way of studying the strategic interactions between players and the overall outcome of their decisions (Myerson 1991). In particular, game theory captures the adversarial nature of the interactions between players (Sinha et al. 2015). In security and defence-related research, game theory is applied to study the behaviour of opposing agents — an attacker and a defender.

2.6 Specific types of hybrid threats, such as cyber threats against cyber-physical systems (Huang & Zhu 2020) or misinformation dissemination in social networks (Kumar & Geethakumari 2013), have been analysed using game theory. More recently, game theory has also been applied to the study of threats that exploit vulnerabilities across domains in a coordinated manner. For example, Keith & Ahner (2021) use game theory to study a scenario in which a defender protects population centres using air defence systems against an attacker capable of both physical and cyberattacks. The attacker seeks to maximise harm by either directly targeting population centres or conducting a cyberattack against the defender's air defence systems. The defender strategically positions their air defence systems to mitigate physical attacks on population centres, and they implement appropriate cyber defences to protect their air defence systems. This research highlights the challenges inherent in multi-domain operations, in which effective defence strategies must account for both physical and cyber threats. Balcaen et al. (2022) use game theory to study the costs associated with hybrid threats. In the game, the attacker can use conventional and unconventional means to destabilise their opponent. The defender decides how to allocate monetary resources to counter such attacks. This research introduces the defence-economic aspects of hybrid threats.

2.7 Such game-theoretic models assume that agents behave rationally and solve for equilibrium. In game theory, agents choose the optimal action to maximise their expected future payoff. There are, however, issues that emerge from this. Mueller (1996) challenges the assumption of pure rationality, suggesting that agents' past experiences significantly influence their decision-making. He argues that individuals often rely on prior knowledge and learned behaviours when making choices, leading to actions that may not always be considered rational. Smith & Price (1973) argue that games that solve for equilibrium are too limited in their representation of real-life problems. The concept of equilibrium, by focusing on a stable outcome in which no player has the incentive to change their strategy, simplifies the complex dynamics of real-life interactions. The concept might not fully account for different factors that influence decision-making, such as emotions, social norms, and evolving relationships between individuals.

2.8 RL, being more dynamic, has the potential to overcome these limitations. RL studies the emergent behaviour of an agent derived from their interactions with a surrounding environment (Sutton & Barto 2020). An RL agent interacts with the environment in discrete time steps. They observe the state of the environment, decide which action to take, and receive rewards for their actions, as well as enter a new environment state. Through trial-and-error interactions, the agent aims to learn an optimal policy that maximises their cumulative reward (Sutton & Barto 2020). Applications of RL include games (Silver et al. 2018), robotics and autonomous systems (Akkaya et al. 2019), recommendation systems (Zheng et al. 2018), information spreading and opinion dynamics in social networks (Banisch & Olbrich 2019; Yu et al. 2016; Hao et al. 2015; Mukherjee et al. 2008), and many more.

2.9 In security and defence science, RL is particularly effective for modelling adversarial behaviours, especially when simulating complex, sequential decision-making processes central to many attack campaigns. For example, an adversary might escalate privileges by gaining initial access, discovering vulnerabilities, and exploiting them to move laterally or gain higher-level access, all while evading detection. RL captures the adaptive nature of these strategies, where each action alters the environment and potentially triggers defensive responses that attackers must learn to bypass. By training RL models in simulated environments, researchers can safely explore and anticipate adversarial actions, providing valuable insights for developing more robust defences.

2.10 Much like game theory, RL methods have mainly been used to study individual threats. Security and defence literature presents research where RL methods have been applied to various aspects of cybersecurity, including intrusion detection (Kurt et al. 2019), intrusion prevention (Hammar & Stadler 2022), and penetration testing (Zhou et al. 2021; Oh et al. 2023; Chowdhary et al. 2020). These problems align well with the RL framework as an agent needs to learn a strategy to optimise an objective through repeated interactions with the surrounding environment. RL methods have also been applied for studying information spreading and opinion dynamics phenomena, such as the emergence of polarisation (Banisch & Olbrich 2019), public opinion expression (Banisch et al. 2022), formation of consensus (Yu et al. 2016) and social norms (Hao et al. 2015; Mukherjee et al. 2008), as well as for detecting and mitigating misinformation (Wang et al. 2020).

2.11 There appear to be no peer-reviewed publications that apply RL to model the behaviour of an adaptive attacker capable of coordinating attacks across cyber and information domains.

Proposed Model

3.1 To devise an agent-based model for further analysing the behaviour of agents with hybrid attack capabilities and their impact on the behaviours and opinions of other agents, we first need to specify the environment's structure, define the types of agents, and describe their behaviour.

Environment

3.2 Common to all RL-based approaches is the need for a training environment. However, the criticality of physical processes prevents direct experimentation with attacks within real-world systems, necessitating the use of virtual testbeds, such as digital twins, to safely simulate and analyse attack-defence scenarios. While existing platforms (Microsoft 2021; DST Group 2022) offer valuable environments for exploring cyberattack and defence operations, they primarily support the study of network-based attacks and lack the ability to represent integrated cyber, physical, and social systems that multi-domain attacks exploit.

3.3 To address this gap, we develop an environment in which it is possible to simulate not only attacks against cyber and physical systems but also misinformation campaigns that can influence peoples' opinions and behaviour. This environment involves a series of competing service providers and their customers. It also includes a social network in which information – both accurate and misleading – about the quality of service can be exchanged. In short, the *environment* represents a Cyber-Physical-Social System (CPSS). Next to physical components and supporting digital elements, a CPSS considers humans and social dynamics as an integral part of the whole system (Wang 2010). Figure 1 visualises the environment.

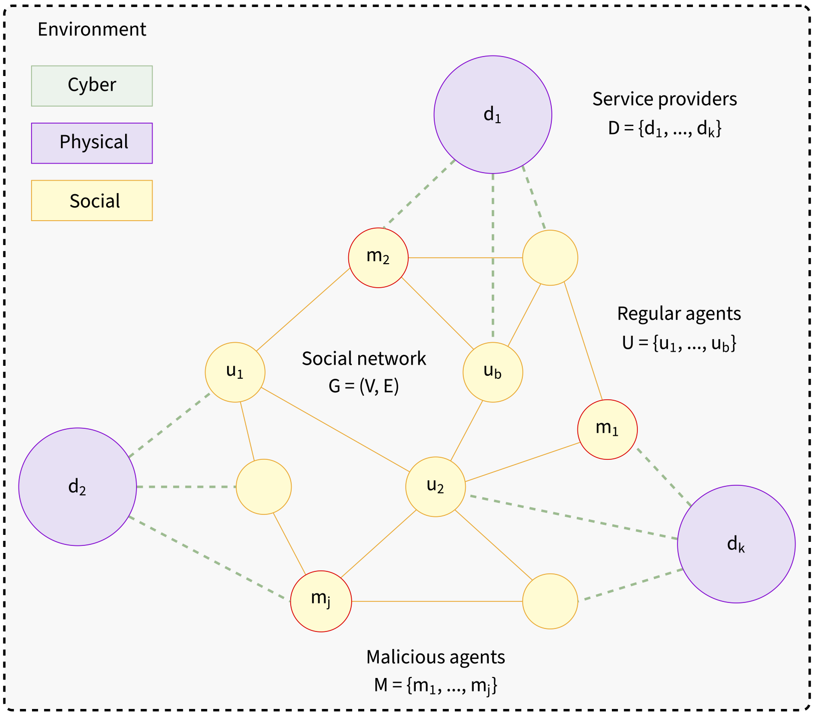

Figure 1: Environment as a Cyber-Physical-Social System.

Cyber-Physical System

3.4 Within the environment, the *Cyber-Physical Systems* are *service providers*, denoted as $D = \{d_1, ..., d_k\}$, with *central* and *end-point states*. This paper assumes that service providers are critical infrastructure companies. For example, a utility company comprises physical components like power generation plants, substations, transmission and distribution lines, and cyber elements like customer portals, communication systems, and smart meters. In this example, a power generation plant responsible for electricity generation has a central state, and the transmission network ending at a customer's premises is the end-point state. To model the possibility of independent failures in central and end-points, we make use of two-state Gilbert-Elliott Markov models (Gilbert 1960; Elliott 1963).

3.5 A service provider, $d \in D$, has a central state, z_d, which can be 1 or 0, indicating whether it is available or unavailable. At each time step t, the new value of the central state is calculated using transition probabilities $\upsilon_d = prob(1 \rightarrow 0)$ and $\psi_d = prob(0 \rightarrow 1)$, as follows:

$$z_{d,t} \leftarrow \begin{cases} 1, & \text{with probability } 1-\upsilon_d, \text{ if } z_{d,t-1}=1 \\ 0, & \text{with probability } \upsilon_d, \quad\;\; \text{if } z_{d,t-1}=1 \\ 1, & \text{with probability } \psi_d, \quad\;\; \text{if } z_{d,t-1}=0 \\ 0, & \text{with probability } 1-\psi_d, \text{ if } z_{d,t-1}=0 \end{cases} \tag{1}$$

3.6 A service provider, $d \in D$, delivers an end-point state, $l_{d,u}$, to each regular agent, $u \in U$. As before, it can either be 1 or 0, showing the availability of the service. At each time step t, the new value of the end-point state for each agent is conditional on the central state and is calculated using transition probabilities $\sigma_d = prob(1 \rightarrow 0)$ and $\theta_d = prob(0 \rightarrow 1)$, as follows:

$$l_{d,u,t} \leftarrow \begin{cases} 1, & \text{with probability } 1-\sigma_d, \text{ if } l_{d,u,t-1}=1 \text{ and } z_{d,t}=1 \\ 0, & \text{with probability } \sigma_d, \quad\;\; \text{if } l_{d,u,t-1}=1 \text{ and } z_{d,t}=1 \\ 1, & \text{with probability } \theta_d, \quad\;\; \text{if } l_{d,u,t-1}=0 \text{ and } z_{d,t}=1 \\ 0, & \text{with probability } 1-\theta_d, \text{ if } l_{d,u,t-1}=0 \text{ and } z_{d,t}=1 \\ 0, & \text{if } z_{d,t}=0 \end{cases} \tag{2}$$

3.7 If the service provider's central state, $z_{d,t}$, is 0, then no end-point service is delivered to their users.

Social System

3.8 Agents, categorised as *regular agents*, $U = \{u_1, ..., u_b\}$, and *malicious agents*, $M = \{m_1, ..., m_j\}$, represent the environment's *Social System*. These agents form a social network described as a static, undirected, and unweighted graph $G = (V, E)$, where an edge $\{v_1, v_2\} \in E$ indicates that agents $v_1, v_2 \in V$ have a social tie between them. The graph is constructed following the Watts & Strogatz (1998) model in which each node initially connects to its κ nearest neighbours in a ring configuration, and each edge is rewired with a probability ρ to a uniformly chosen random vertex. This model generates a social network with short average path lengths between nodes and high clustering, which are two inherent characteristics of social networks (Watts & Strogatz 1998). Agents can exchange information with each other through the social network and interact with the service providers through online communication systems. By combining agents' social interactions with service providers' cyber-physical components, the environment represents a CPSS.

Agent-based model

3.9 Our agent-based model details the behaviour of malicious and regular agents in the environment. We represent the behaviour of malicious agents with an attack model and the behaviour of regular agents with an experience model and an opinion dynamics model. The model's notation is added as a table in Appendix A.

Attack model

3.10 An attack model describes how malicious agents learn about the system's dynamics and use their experience and collected information to attack their target. Whilst not a limitation of the approach, this paper assumes that malicious agents can launch two different types of attacks: (1) a denial of service attack against a service provider that blocks the availability of targeted resources to legitimate users and (2) a misinformation campaign in a social network that manipulates public opinion. Malicious agents can coordinate these attack methods to exploit vulnerabilities across cyber and information domains. Their goal is to maximise the number of service provider customers who experience attacks by directly experiencing service unavailability or through exposure to negative information on the social network, both of which may cause customers to change their provider.

3.11 The impact of the attacks is expected to depend on their timings and coordination. The first key question is *when* to attack to realise the attacker's goals. Whether an attacker attacks too early or too late, an opportunity to have a more significant impact might be lost. The second question is for *how long* to attack a system. The longer the time for which malicious agents attack a system at a given level of intensity, the greater the probability that the attack is detected. Therefore, an attacker aims to achieve their goal as quickly as possible or as quietly as possible to reduce the chance of detection. The third question is in *which way should attacks be combined* to potentially maximise harm to the target. A misinformation campaign can sow distrust that the adversaries can later exploit with a cyberattack. Alternatively, a cyberattack could first disrupt critical systems, creating an opportunity for the attackers to disseminate tailored misinformation to amplify harm.

3.12 Malicious agents observe the *state* of the environment, $s \in \mathcal{S}$, that indicates the number of regular agents associated with each service provider. This number represents the agents an attack campaign can impact. They have 5 available *actions* $a_m \in \mathcal{A}_m = \{0, 1, ..., 4\}$, in which action 0 means undertaking reconnaissance, $1 =$ conducting a cyberattack, $2 =$ spreading misinformation, $3 =$ combining attacks (cyberattack and misinformation), and $4 =$ terminating the attack campaign. Malicious agents receive a reward, r_m, for their actions. This reward consists of the following components:

- The cyberattack reward measures the proportion of regular agents affected by an attack against a service provider. It is negative during reconnaissance (indicating the cost of inaction) and positive after attack initiation.
- The misinformation reward measures the proportion of regular agents exposed to misinformation spread by malicious agents. It is negative during reconnaissance (indicating the cost of inaction) and positive after attack initiation.
- The detection reward represents the proportion of regular agents that have detected either a cyberattack or misinformation. It is initially zero and negative after attack initiation.

- regular agents who request service from the affected provider have a probability of detecting a cyberattack controlled by the parameter ζ.
- regular agents in contact with malicious agents spreading misinformation have a probability of detecting misinformation controlled by the parameter η.

3.13 Malicious agents start in the reconnaissance stage. We assign a positive value ξ to undertaking reconnaissance action, encouraging malicious agents to observe the environment. At each time step t, they receive the state of the environment, $s_t \in S$, and take action following their policy π_m as:

$$\pi_m = \operatorname*{arg\,max}_{a \in \mathcal{A}_m} Q_m(a) \tag{3}$$

where $Q_m(a)$ is an action-value function. When staying in reconnaissance is no longer beneficial, they target the service provider whose disruption could potentially have the most far-reaching consequences. This decision is based on the potential to inflict maximum harm on customers through service outages and the propagation of misleading information. For every action, they receive a reward, which is used to update the action-value, $Q_m(a)$, as follows:

$$Q_m(a) \leftarrow Q_m(a) + \alpha \times [r_m - Q_m(a)] \tag{4}$$

where a is the selected action, r_m is the received reward, and α is the learning rate. Figure 2 illustrates the behaviour of malicious agents in our environment.

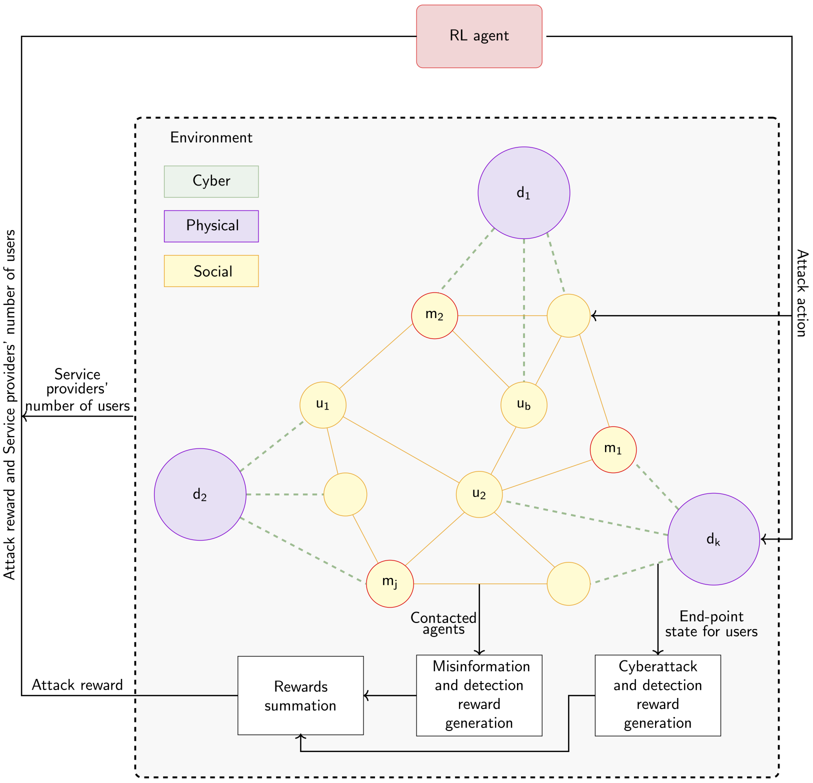
RL agent
Environment
Cyber
Physical
Social
d$_1$
m$_2$
u$_1$
u$_b$
m$_1$
d$_2$
u$_2$
d$_k$
m$_j$
Attack action
Service providers' number of users
Attack reward and Service providers' number of users
Contacted agents
End-point state for users
Rewards summation
Misinformation and detection reward generation
Cyberattack and detection reward generation
Attack reward

Figure 2: Malicious agents' behaviour in the environment.

Experience model

3.14 Each regular agent has a service provider and several social connections, as illustrated on Figure 1. To choose service providers, agents combine their personal experiences with information from their social network.

3.15 Each agent, $u \in U$, follows a policy, π_u, that balances their direct experiences of using the service (represented by action values $Q_u(a)$) with information about providers heard from their social connections (represented by opinion values $\Phi_u(o)$). The agent's policy is an ϵ-greedy policy defined as follows:

$$\pi_u = \begin{cases} \operatorname*{arg\,max}_{a \in \mathcal{A}_u} \left\{ \omega \times Q_u(a) + (1-\omega) \times [\Phi_u(o_a^+) - \Phi_u(o_a^-)] \right\} & \text{with probability } 1-\epsilon \\ \mathcal{U}(\mathcal{A}_u) & \text{with probability } \epsilon \end{cases} \tag{5}$$

where $\Phi_u(o_a^+)$ and $\Phi_u(o_a^-)$ represent the agent's positive and negative opinions about action a, ω is a weight controlling the relative importance of direct experiences versus social information, ϵ is an exploration parameter, and $\mathcal{U}(\mathcal{A}_u)$ denotes a sample from the uniform distribution over the action space. By considering the difference between positive and negative opinion values next to action values, agents are more likely to choose providers about whom they have a stronger positive sentiment or avoid those with negative reputations.

3.16 Regular agents have several available actions representing the choice between service providers. At each time step t, each agent, $u \in U$, takes action, $a_{u,t} \in \mathcal{A}_u = \{d_1, ..., d_k\}$, and receives service from their chosen provider, d_k. They record the received end-point state, $l_{d,u,t}$, as their reward value, r_u. During the learning process, they update their evaluation of actions, $Q_u(a)$, as:

$$Q_u(a) \leftarrow Q_u(a) + \alpha \times [r_u - Q_u(a)] \tag{6}$$

where a is the selected action, r_u is the received reward value, and α is a learning rate. Each agent then asks for information from one of their neighbours, $c_u \in C_u$. This process is explained in *Opinion dynamics model* section. After receiving information from their neighbour, the agent continues following the ϵ-greedy policy to select the next action.

Opinion dynamics model

3.17 An opinion dynamics model describes how regular agents interact with their neighbours and use the received information to develop their opinions on service providers' behaviour. This paper follows the examples of Banisch & Olbrich (2019) and Yu et al. (2016) for designing opinion-sharing and adoption mechanisms.

3.18 Each agent, $u \in U$, follows a policy, μ_u, represented by a vector of opinion values, $\Phi_u(o)$. The policy of the agent is an ϵ-greedy policy defined as:

$$\mu_u = \begin{cases} \text{arg max}_{o \in \mathcal{O}_u} \Phi_u(o) & \text{with probability } 1-\epsilon \\ \mathcal{U}(\mathcal{O}_u) & \text{with probability } \epsilon \end{cases} \tag{7}$$

where o is an opinion, ϵ is an exploration parameter, and $\mathcal{U}(\mathcal{O}_u)$ denotes a sample from the uniform distribution of the opinion space.

3.19 Regular agents have $2 \times k$ available opinions. Each opinion, $o_u \in \mathcal{O}_u = \{o_1^-, o_1^+, ..., o_k^-, o_k^+\}$, reflects the agent's sentiment towards a service provider's behaviour. Each opinion is a tuple of an opinion value and a service provider's identity. The opinion value is 1 or -1, indicating a positive or negative opinion. For example, a positive opinion on service provider d is a tuple of $(1, d) = o_d^+$. Similarly, a negative opinion is $(-1, d) = o_d^-$.

3.20 At each time step t, each agent, $u \in U$, expresses an opinion, $o_{u,t} \in \mathcal{O}_u$, to a randomly chosen neighbour, $c_u \in C_u$, in the social network. This way, agents are exposed to a wider range of opinions that can help them to make informed decisions. When asked, the neighbour compares the proportion of service availability, $\overline{l_{d,u}}$, calculated as $\overline{l_{d,u}} = \sum_{t=1}^{t}(l_{d,u,t} \times \delta_{a_{u,t},d})$, where $d = o_{u,t}^{(2)}$, with a satisfaction threshold, τ, to decide whether they have a positive or negative opinion of the specific service provider. If the neighbour has no prior experience with the service provider, they do not provide feedback. Otherwise, if their opinions match, they get a positive reward, $h_u = 1$; alternatively, they get a negative reward of -1. They update their internal evaluation of opinion, $\Phi_u(o)$, as:

$$\Phi_u(o) \leftarrow \Phi_u(o) + \alpha \times [h_u(o) - \Phi_u(o)] \tag{8}$$

where o is the expressed opinion, h_u is the reward value, and α is a learning rate. Afterwards, they continue following the ϵ-greedy policy for selecting an action and expressing an opinion at the next step. Figure 3 illustrates the behaviour of each regular agent as an RL agent in our custom environment.

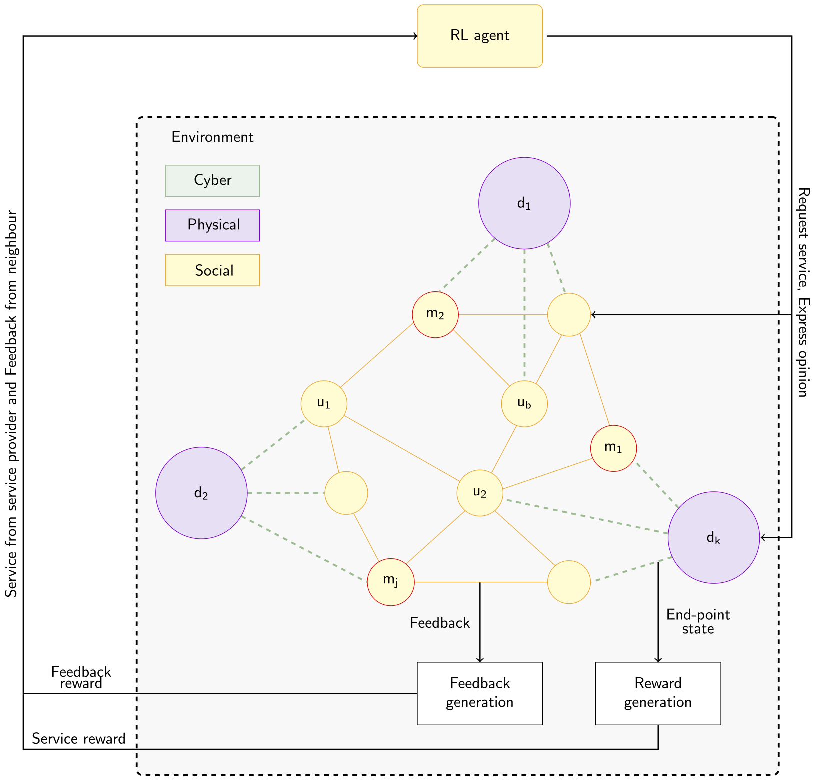

Figure 3: Regular agents' behaviour in the environment.

Parameters and input values

3.21 Table 1 summarises the parameters in the simulation. These values were chosen on the basis of a combination of theoretical considerations and practical limitations:

- The learning rate, α, set to 0.05, controls how quickly agents learn from their experiences. A smaller value ensures stability during learning, while a larger value might lead to faster but potentially less robust learning.

- The exploration parameter, ϵ, set to 0.10, balances the exploitation of known good choices with the exploration of new service providers and opinions.

- The cyberattack and misinformation detection probabilities, ζ and η, set to 0.80 and 0.10, reflect the assumed relative difficulty in detecting these attacks. These values reflect the assumption that people are more likely to recognise a cyberattack than a well-crafted misinformation campaign.

- The number of neighbours, κ, is set to 6, following values used in social network studies (Watts & Strogatz 1998; Perra & Rocha 2019). The graph was generated using the Watts & Strogatz (1998) model with a rewiring probability, ρ of 0.01. This probability value introduces a small degree of randomness into the network structure while primarily preserving local connections, reflecting realistic social network patterns.

- The satisfaction threshold, τ, set to 0.80, represents the expected level of service availability that regular agents consider satisfactory.

- The direct experience weight, ω, set to 0.80, indicates how much agents value their personal experiences over information from their social network. This value reflects the assumption that agents generally trust their own experiences more than information from others.

- The transition probability values, υ, ψ, σ, and θ, are hard to estimate due to limited data; therefore, the probability values, shown in Gilbert (1960), are used to model the output of a service provider's end-point state. The transition probabilities are chosen so that the central state is more likely to be available than the end-point states, reflecting greater stability and availability. In our example, this translates to a power generation plant being less prone to failure than a transmission network.

Parameter	Description	Value
α	Learning rate	0.05
ϵ	Exploration parameter	0.10
ζ	Cyberattack detection probability	0.80
η	Misinformation detection probability	0.10
κ	Number of neighbours	6
ρ	Rewiring probability	0.01
τ	Satisfaction threshold	0.80
ω	Direct experience weight	0.80
ξ	Positive initialisation of malicious agents	100
σ	End-point transition probability $1 \rightarrow 0$	0.03
θ	End-point transition probability $0 \rightarrow 1$	0.25
υ	Central state transition probability $1 \rightarrow 0$	0.01
ψ	Central state transition probability $0 \rightarrow 1$	0.30

Table 1: Parameter values in the model.

3.22 Table 2 summarises input values used in the simulation. Warm-up time, W, necessary for different aspects of the simulation to normalise, is the period for which the simulation will run before any data is collected. Welch (1951) method was used to determine the length of the warm-up time. Observations showed convergence of the simulation after 2000 time steps. The number of time steps, T, set to 10000, is proportional to the warm-up time. The simulation environment includes 1000 agents, of which 950 are regular agents and 50 (5%) are malicious agents. This setup allows the model to focus on the behaviour of the majority of agents while examining the disruptive influence of a minority of malicious actors. The three service providers introduce competition into the system, creating options for regular agents. The chosen agent population size balances computational efficiency with the ability to observe interesting social network dynamics.

Input	Description	Value
T	Number of time steps	10000
W	Warm-up time	2000
$b, \|U\|$	Number of regular agents	950
$j, \|M\|$	Number of malicious agents	50
$k, \|D\|$	Number of service providers	3

Table 2: Input values in the model.

Experiments and Analysis

4.1 The section presents metrics used to characterise agents' behaviour and introduces the results of two experiments. The first experiment concerns the behaviour of regular agents alone, i.e., without introducing adversarial events. The second experiment examines the attackers' strategic behaviour and impact on regular agents. We conducted a linear regression analysis to understand which malicious agents' decisions impacted regular agents' behaviour and opinions.

Metrics

4.2 This section introduces metrics that were used to understand the system's behaviour.

1. Average service selection rate, $\overline{n_d}$, shows the regular agents' preference for each service provider. Higher values indicate a preference among these agents for a particular service provider. We calculate this as follows:

$$\overline{n_d} = \frac{1}{|U| \times T} \times \sum_{u=1}^{|U|} \sum_{t=1}^{T} \delta_{a_{u,t},d} \ \mid a = \text{choose } d \tag{9}$$

where $\delta_{x,y}$ is the Kronecker delta, equal to 1 if $x = y$ and 0 otherwise, action a is the choice of service provider d, $|U|$ is the number of users, and T is the number of time steps.

2. Average service level, $\overline{l_d}$, is the proportion of time service was available when requested. Higher values indicate greater service availability. We calculate this as follows:

$$\overline{l_d} = \frac{1}{|U| \times T} \times \sum_{u=1}^{|U|} \sum_{t=1}^{T} (l_{d,u,t} \times \delta_{a_{u,t},d}) \tag{10}$$

   where $l_{d,u,t}$ is the end-point state provided by service provider d to agent u at time step t, $\delta_{x,y}$ is the Kronecker delta, equal to 1 if $x = y$ and 0 otherwise, action a is agent u's choice of service provider d, $|U|$ is the number of users, and T is the number of time steps.

3. Average opinion value on service provider, $\overline{o_d}$, shows the strength of positive o_d^+ or negative opinion o_d^- on service provider, calculated as follows:

$$\overline{o_d} = \frac{1}{|U| \times T} \times \sum_{u=1}^{|U|} \sum_{t=1}^{T} \Phi_{u,t}(o) \tag{11}$$

   where $\Phi_{u,t}(o)$ is agent u's value of opinion o at time step t, $|U|$ is the number of users, and T is the number of time steps.

4. Average regret, $\overline{\nu}$, is a performance metric that shows the difference between the optimal and received rewards. A smaller value indicates that regular agents, on average, are making decisions closer to the optimal strategy. We calculate this as follows:

$$\overline{\nu} = \frac{1}{|U| \times T} \times \sum_{u=1}^{|U|} \sum_{t=1}^{T} (r_t^* - r_{u,t}) \tag{12}$$

   where r_t^* is the optimal reward value at time step t ($r_t^* = 1$ when any service provider's end-point state $l_{d,u,t} = 1$ and $r_t^* = 0$ otherwise), $r_{u,t}$ is the reward received by user u at time step t, $|U|$ is the number of users, and T is the number of time steps.

5. Attack duration is the number of time steps for which an attack took place.
   - Cyberattack duration, x_1, is the number of time steps for which the attack took place.
   - Misinformation duration, x_3, is the number of time steps for which a misinformation campaign took place.
   - Combined attack duration, x_5, is the number of time steps during which an attacker used both attack methods at the same time.

6. Attack start marks the time step at which an attack started.
   - Cyberattack start, x_2, is the time step at which malicious agents launch a DoS attack against a service provider.
   - Misinformation start, x_4, is the time step at which malicious agents launch a misinformation campaign in the social network.
   - Combined attack start, x_6, is the time step at which an attacker decides to start using both attack methods simultaneously.

7. Order of attacks, x_7, shows the sequence in which malicious agents launch attacks. It is a binary variable recorded as 1 if malicious agents start their attack campaign with a cyberattack and launch a misinformation campaign later. Alternatively, if they start with misinformation and escalate the situation with a cyberattack, the variable is recorded as 0.

8. Impact on service selection rate, y_1, is the difference between the average service selection rate during an attack and during normal operations (no attack).

9. Impact on positive opinion, y_2, is the difference between the average positive opinion value during an attack and during normal operations (no attack).

10. Impact on negative opinion, y_3, is the difference between the average negative opinion value during an attack and during normal operations (no attack).

Experiment 1

4.3 In the first experiment, we investigated the general behaviour of regular agents in the environment. For this, we ran the agent-based model without introducing malicious behaviour. We measured each service provider's average service selection rate, average service level, and average opinion values to understand how regular agents and service providers behave in the environment. We also measured the average regret value to understand how well agents performed. Table 3 summarises the results.

Metric	Description	Mean	Std Dev	Min	Max
$\overline{n_1}$	Average service selection rate	0.33	0.03	0.22	0.42
$\overline{l_1}$	Average service level	0.89	0.00	0.89	0.91
$\overline{o_1^+}$	Average positive opinion	0.94	0.06	0.63	1.00
$\overline{o_1^-}$	Average negative opinion	-0.94	0.06	-0.64	-0.99
$\overline{n_2}$	Average service selection rate	0.33	0.03	0.21	0.41
$\overline{l_2}$	Average service level	0.89	0.00	0.88	0.91
$\overline{o_2^+}$	Average positive opinion	0.96	0.06	0.57	1.00
$\overline{o_2^-}$	Average negative opinion	-0.96	0.06	-0.57	-1.00
$\overline{n_3}$	Average service selection rate	0.34	0.02	0.27	0.39
$\overline{l_3}$	Average service level	0.89	0.00	0.88	0.90
$\overline{o_3^+}$	Average positive opinion	0.97	0.03	0.82	1.00
$\overline{o_3^-}$	Average negative opinion	-0.97	0.03	-0.82	-0.99
$\overline{\nu}$	Average regret	0.10	0.00	0.10	0.11

Table 3: Agents behaviour in the first experiment. The number of simulation runs was 100. The index of a metric identifies the service provider.

4.4 We found that all service providers had similar popularity among regular agents, each providing service to about a third (between $33 - 34\%$) of the agents. Decisions on which service provider to use were based on what regular agents directly experienced and what they heard from others in the social network. As service providers' transition probabilities between states, $\sigma, \theta, \upsilon, \psi$, that indicate service availability, were the same (see Table 1), their average service level (0.89) was also similar. The average service level exceeded the satisfaction threshold τ, making regular agents form positive opinions of each service provider. As regular agents generally had good experiences and thought well of all the providers, they chose between them fairly equally. Regular agents formed positive opinions of each service provider's behaviour (average values between 0.94 and 0.97) based on the feedback they received from the agents in their social network. These values imply that users strongly believed that all service providers behaved well. Positive values for positive opinions show that when users asked their neighbours if they were satisfied with their service provider, their neighbours tended to agree. Negative values for negative opinions (average values between -0.94 and -0.97) show that when regular agents asked their neighbours if they were unsatisfied, they disagreed. We expected such results as each service provider exceeded the satisfaction threshold τ with their performance. We measured agents' average regret value as 0.10, which aligns with the exploration parameter ϵ. This fixed exploration parameter is a primary limiting factor in the agent's ability to achieve optimal decisions consistently. To further reduce the regret value, it would be necessary to decrease the exploration parameter value during training dynamically.

4.5 These results show what regular agents learned to do in the environment to maximise their cumulative reward while balancing exploitation with exploration. Because their transition probabilities were the same, we expected minor differences between the average service levels of the three service providers. In addition, as the average service levels were higher than the satisfaction threshold, we expected that users formed positive opinions on the service providers' performance. With this experiment, we showed how the dynamics in the system evolved without malicious activities.

Experiment 2

4.6 In this experiment, we investigated the behavioural strategies of malicious agents with coordinated attack capabilities and their impact on others' behaviours and opinions. We included both regular and malicious agents in the environment. We first measured the average service selection rate, average positive and negative opinion values, average service level, and regret to gain insight into the changes in agents' behaviour. We then conducted a linear regression analysis on simulated data to understand the impact. Table 4 summarises the results of the second experiment.

Metric	Definition	Mean	Std Dev	Min	Max
$\overline{n_1}$	Average service selection rate	0.31	0.06	0.11	0.42
$\overline{l_1}$	Average service level	0.88	0.00	0.86	0.89
$\overline{o_1^+}$	Average positive opinion	0.86	0.11	0.37	0.96
$\overline{o_1^-}$	Average negative opinion	-0.85	0.10	-0.38	-0.94
$\overline{n_2}$	Average service selection rate	0.35	0.06	0.15	0.51
$\overline{l_2}$	Average service level	0.88	0.01	0.86	0.89
$\overline{o_2^+}$	Average positive opinion	0.92	0.07	0.53	1.00
$\overline{o_2^-}$	Average negative opinion	-0.90	0.07	-0.53	-0.96
$\overline{n_3}$	Average service selection rate	0.34	0.06	0.19	0.48
$\overline{l_3}$	Average service level	0.88	0.01	0.86	0.89
$\overline{o_3^+}$	Average positive opinion	0.91	0.08	0.58	0.99
$\overline{o_3^-}$	Average negative opinion	-0.90	0.08	-0.58	-0.96
$\overline{\nu}$	Average regret	0.12	0.00	0.11	0.13

Table 4: Agents behaviour in the second experiment. The number of simulation runs was 100. The index of a metric identifies the service provider.

4.7 We observed that the average service selection rates were between 31% and 35%, similar to the first experiment's findings. However, the average service level was slightly lower because, in each simulation, malicious agents targeted a service provider, making it unavailable to its users. These attacks resulted in a decrease in the average service level. We also noticed differences in the average positive and negative opinion values compared to the first experiment. For instance, the average positive opinion on the first service provider was smaller (0.86 compared with 0.94), indicating that agents in this experiment believed less in the positive opinion of the service provider to be true. The average negative opinion values on service providers were also smaller (e.g., -0.85 compared with -0.94), indicating that agents were less confident that a negative opinion on a service provider was false. There was a difference in regret values between the experiments; 0.12 compared to 0.10 suggested that agents performed worse following their policy in this experiment compared to the first experiment.

Impact analysis

4.8 We conducted regression analyses to identify which malicious agents' strategic decisions impacted regular agents' behaviour and opinions in the environment. For this, we measured the start time of an attack, its duration, the order of attacks, the impact on service selection rate, the impact on positive opinion value, and the impact on negative opinion value.

4.9 We found that attack start time variables had a low correlation (value < 0.20) with all impact variables. Therefore, these variables (i.e., x_2, x_4, and x_6) were excluded from further analysis. As combined attack length can be calculated from other attack variables, its relationships with impact variables were observed separately. The data showed that malicious agents always started their attack campaign with a cyberattack and launched a misinformation campaign later. Therefore, as there was no data on the results of an alternative order of attacks, we had to exclude the order of attacks from further analysis. Combining the remaining independent variables (x) with dependent variables (y), introduced in Section 4.2, we ended up with 6 different models to analyse. Table 5 shows these results.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
b_0	-0.090 * (0.04)	-0.053 * (0.02)	-0.049 (0.03)	0.046 * (0.02)	0.005 (0.03)	0.081 *** (0.02)
x_1	0.016 *** (0.00)		0.004 * (0.00)		0.003 * (0.00)	
x_3	0.002 (0.00)		0.007 *** (0.00)		0.005 *** (0.00)	
x_5		0.016 *** (0.00)		0.005 ** (0.00)		0.005 *** (0.00)
R^2	0.48	0.48	0.19	0.09	0.20	0.11
Adj. R^2	0.47	0.47	0.18	0.08	0.19	0.10

Table 5: Regression analysis results. The number of simulation runs was 100. Standard errors in parentheses. *** $p \leq 0.001$; ** $p \leq 0.01$; * $p \leq 0.05$; † $p \leq 0.1$

4.10 The first model *(Model 1)* predicts the impact on service selection rate given the length of a cyberattack x_1 and misinformation campaign x_3. The following equation predicts the impact on the service selection rate as:

$$y_1 = -0.090 + 0.016 \times x_1 + 0.002 \times x_3 + \epsilon \tag{13}$$

where ϵ is the error term. The model results showed that the impact on service selection rate depends significantly on the length of a cyberattack but does not depend significantly on the length of a misinformation campaign. The model suggests that more regular agents tend to change their service provider the longer it is unavailable due to a cyberattack. According to the model, an attacker needs to find resources to conduct a cyberattack for at least 6 time steps to receive a significant impact on the user behaviour. If attackers decide to attack a service provider for an additional time step, the impact on the service selection rate will increase by 0.016 units. Here, we found that around 1.6% of regular agents will leave their service provider during every additional time step of a cyberattack. Analysis of R^2 showed that the attack length variables collectively explained about 48% of the variability observed in the impact variable.

4.11 The second model *(Model 2)* predicts the impact on service selection rate given the length of a combined attack. The equation for this is:

$$y_1 = -0.053 + 0.016 \times x_5 + \epsilon. \tag{14}$$

where ϵ is the error term. The results showed that an attacker needs to conduct a cyberattack with a misinformation campaign for at least 4 steps to achieve a positive impact. The longer the combined attack lasts, the higher the positive impact is for malicious agents. Here, we found that the percentage of regular agents who will leave their service provider is 1.6% for every additional time step of a combined attack. Like the first model, the combined attack duration variable explains around 48% of the variability in the dependent variable.

4.12 The third model *(Model 3)* predicts the impact on the positive opinion given the lengths of a cyberattack and misinformation campaign. The equation is:

$$y_2 = -0.049 + 0.004 \times x_1 + 0.007 \times x_3 + \epsilon \tag{15}$$

where ϵ is the error term. The results showed that the impact on the positive opinion value significantly depends on both independent variables. While spreading misinformation for only 1 time step, an attacker should attack a service provider for at least 23 time steps to influence the positive opinion value. However, the longer they conduct a cyberattack, the more likely they will be detected. Misinformation detection can be more challenging. According to the model, they only need to spread misinformation for 7 time steps and a cyberattack for 1 time step to influence the positive opinion value. If an attacker decides to attack the service provider for 1 time step longer, the impact will increase by 0.004 units. If an attacker decides to spread misinformation about a service provider in the social network for 1 time step longer, the impact will increase by 0.007. Opinion values show the strength of an opinion in society. Here, we found that the strength of a positive opinion value of the targeted service provider will decrease by either 0.004 or 0.007 units. Considering their resources, attackers can decide how long to conduct a cyberattack or a misinformation campaign to change the positive opinion of the targeted provider. The R^2 value showed that the attack length variables explained only 19% of the variability in the impact variable.

4.13 The fourth model *(Model 4)* predicts the impact on the positive opinion given the length of a combined attack as:

$$y_2 = 0.046 + 0.005 \times x_5 + \epsilon \tag{16}$$

where ϵ is the error term. The results showed that the impact on positive opinion value significantly depends on the length of a combined attack. The results suggested that conducting a combined attack for 1 time step longer will increase the impact by 0.005 units. The strength of the positive opinion value on the targeted service provider will decrease by 0.005 at every additional time step. The R^2 value showed that the independent variable explained only around 9% of the variability in the dependent variable.

4.14 The fifth model (*Model 5*) predicts the impact on the negative opinion given the length of a cyberattack and misinformation campaign. The following equation predicts the impact variables as:

$$y_3 = 0.005 + 0.003 \times x_1 + 0.005 \times x_3 + \epsilon \tag{17}$$

where ϵ is the error term. The results showed that the dependent variable significantly depends on the independent variables. According to the model, increasing the cyberattack length variable by one additional time step increases the impact on the negative opinion value by 0.003 units. Similarly, increasing the number of misinformation campaign time steps by 1 increases the impact on the negative opinion value by 0.005. Here, we found that the length of a misinformation campaign increased the impact on the negative opinion value more than a cyberattack did. We expected this as we interpreted "misinformation" as saying that the service provider's performance was negative even when it was not true. The independent variables explain about 20% of the variability in the dependent variable.

4.15 With the last model (*Model 6*), we wanted to predict the impact on the negative opinion value given the length of the combined attack. The equation expresses this as:

$$y_3 = 0.081 + 0.005 \times x_5 + \epsilon \tag{18}$$

where ϵ is the error term. The results showed that the impact on negative opinion value significantly depended on the combined attack length. Here, we found that increasing the length of a combined attack by 1 time step will increase the impact on negative opinion by 0.005. The results showed that a combined attack would increase the strength of the negative opinion in society. The R^2 values showed that the combined attack length variable explained around 11% of the variability in the negative opinion value. The R^2 values from this and previous models offer valuable insights. They also highlight the opportunity to explore a potentially richer set of factors influencing agents' behaviour and opinions in real scenarios.

4.16 We tested the linear regression assumptions for all models before analysing the results. We confirmed that the relationship between the predictor and outcome variables was linear; the residuals were normally distributed, uncorrelated, and had a constant variance. Sensitivity analysis results are added as Appendix B.

Discussion

5.1 The threat landscape continues to evolve rapidly, with attackers using more advanced tools, like artificial intelligence (AI) and machine learning (ML), to launch large-scale, sophisticated attacks at lower cost (Brundage et al. 2018; Bresniker et al. 2019). Large Language Models, for example, can be used to generate large volumes of misleading or false information, including personalised dialogue, which can be used for targeted disinformation campaigns. Attackers can also use AI/ML techniques to develop attacks against specific targets. By creating a digital twin as a digital copy of a targeted system, they can simulate attacks, assess potential harm, and refine their tactics before launching the attack against a real-world system. The prospect of such autonomous and adaptive attacks raises concerns for the future of defence.

5.2 Traditional security analysis approaches often rely on game theory to study attacker behaviour. However, hybrid threats, characterised by adaptive behaviour and complex attack strategies, can undermine the core rationality assumption and equilibrium concept that underlie game theory. Furthermore, the limited real-world data on hybrid threats and their consequences make it difficult to develop accurate and reliable data-driven models. These issues create a pressing need for analytical tools and methods that can adapt and learn, even in situations in which information is limited. Such tools should be capable of representing complex multi-domain attack situations and enabling a defence-in-depth approach that addresses technical and social issues.

5.3 We address this problem by proposing a novel agent-based model in which malicious agents use RL for decision-making. Instead of having to pre-define every attacker movement in the environment, the approach allows them to learn and improve their attack strategies based on their experiences over time. Future work can expand upon this foundation by incorporating additional factors, such as resource constraints and the presence of autonomous defence mechanisms, into the malicious agents' decision-making processes. This will advance the understanding of how attackers adapt their behaviour in response to evolving defences, ultimately leading to the development of more robust defence strategies.

5.4 We determined the effectiveness of our RL approach by analysing how the attackers' decisions on attack start time, duration, and coordination of attacks affected the societal impact. We found a weak correlation between attack timing and societal impact, contradicting prior research that showed the potential of an attack during system downtime to cause considerable harm (Krotofil et al. 2014). Our result suggests that attackers with hybrid attack capabilities using RL may struggle to consistently identify these optimal attack moments. The reason for this can be the forward-looking nature of attack planning; attackers must weigh current opportunities against potentially better future scenarios. The decision-making process becomes even more complex when attackers operate across multiple interconnected domains. They face the additional challenge of predicting the cascading effects of coordinated attacks across domains. Accurately predicting these interactions is hard.

5.5 Our findings on attack campaign duration and societal impact align with prior research, highlighting a correlation between longer durations and greater societal impact (Podobnik et al. 2015). Attackers with hybrid attack capabilities face the complex problem of determining the balance between the acceptable cost and the received benefit of their attack campaign. Extending the duration can indeed increase the potential societal harm by allowing the effects to cascade across domains. At the same time, they need to balance potential gains against the increased resource expenditure required to sustain the attack across multiple domains and the heightened likelihood of detection due to the extended timeframe.

5.6 Our investigation into attack strategies revealed that attackers consistently initiated campaigns with cyberattacks followed by misinformation. The initial cyberattack disrupts critical systems, potentially leading to operational challenges and an increased susceptibility to further manipulations. The subsequent misinformation campaign then takes advantage of this situation by influencing perceptions or shaping narratives. Following this strategy, the attacker can potentially reduce misinformation detection as it is likely to blend with the more genuine concerns caused by the cyberattack. Additionally, after a cyberattack, people are likely more susceptible to misinformation, amplifying its societal impact.

5.7 While our research suggests limitations to attackers' ability to optimise all aspects of attacks (timing, duration, coordination of methods) using RL, it highlights the broader threat of intelligent attackers leveraging advanced technologies. Autonomous penetration testing tools, used for defence purposes today (Oh et al. 2023; Zhou et al. 2021; Chowdhary et al. 2020), demonstrate the feasibility of automated vulnerability detection and exploitation. These tools have the potential to be adapted for malicious purposes, generating and executing autonomous attack plans across diverse environments (Chowdhary et al. 2020). The lack of current evidence about attackers actively using such tools in real life should not create a false sense of security. The increasing availability of powerful computing resources, digital twins, and huge datasets will likely accelerate the development of more sophisticated malicious AI capabilities in the future.

5.8 While RL has proven to be a valuable component in detecting and mitigating misinformation within social networks, as shown in Wang et al. (2020), the same techniques raise concerns regarding their potential malicious use. It could be used for tailoring misinformation campaigns or testing the effectiveness of various narratives on different groups based on their preferences and online behaviours. Additionally, RL could counter defensive measures against misinformation spreading by continually learning and adapting to evade detection and maximise impact.

5.9 Our study expanded upon these independent threats by exploring the potential for intelligent attackers to generate autonomous attack operations across multiple domains. While our focus was only on the attacker's behaviour, real-world scenarios would likely include defensive countermeasures. This underscores the importance of modelling hybrid threats and defensive countermeasures as an adversarial multi-agent attack and defence problem, where both agents employ evolving, intelligent, tactics. Just as autonomous attacks are plausible, so too is the potential for autonomous, adaptive defence.

Conclusion

6.1 We proposed an agent-based model composed of intelligent learning agents with which to study the strategic behaviour of malicious agents with hybrid attack capabilities. We implemented the proposed model and performed linear regression analysis on the collected data. With the model, we were able to determine the strategic behaviour of malicious agents that would potentially maximise their influence on the target audience.

6.2 In summary, our research showed that malicious agents with hybrid attack capabilities could effectively learn to attack a system. They learned to decide the length of an attack on the system so that agents would start changing their behaviour and opinions about the target. The longer the attack lasted, the higher the impact was, but the greater the chance of detection. They also learned to exploit vulnerabilities across cyber and information domains in a coordinated manner that significantly impacted the target. The results showed that an adversary could effectively learn to coordinate attacks across cyber and online information domains based on their understanding of the system's behaviour. Such adaptive behaviour can be particularly challenging for other agents to detect. We also conducted a sensitivity analysis of the model's parameters to determine how different parameter values affect the outcome variables given the set of assumptions.

6.3 RL holds significant potential for modelling complex attack scenarios that require adaptive strategies, precise timing, and coordinated efforts. An RL agent's ability to learn and adapt continuously through interaction with its environment allows it to craft sophisticated strategies that exploit specific vulnerabilities in defences, such as identifying unpatched vulnerabilities or intensifying resources during an attack to overwhelm defenders' responses. Multi-agent RL further extends this capability by simulating scenarios in which multiple adversarial agents must coordinate actions, such as a synchronised cyber and misinformation campaign, to achieve shared objectives. Additionally, multi-agent RL can be used in adversarial environments where both attackers and defenders are learning agents, dynamically adjusting their strategies in response to each other. This dynamic interplay between agents offers valuable insights into complex adversarial interactions, highlighting RL's distinct advantages for simulating advanced attack strategies. These capabilities lay the groundwork for future research, expanding RL's application in attack modelling and helping to develop more robust defensive countermeasures.

6.4 Future work will focus on advancing malicious agents' learning behaviour to improve their decision-making on the optimal time to attack the system. We will analyse in which order malicious agents should learn to coordinate attacks to maximise their impact on the behaviours and opinions of other agents in the system. We will evaluate different scenarios to determine the most impactful strategies, and will explore the differences in impact between individual attacks and coordinated attacks. Future work will also analyse how to respond effectively to such attacks. More specifically, we will develop and implement relevant detection and response mechanisms to limit the impact of coordinated attacks in multi-agent systems.

Model Documentation

The model is available at: `https://www.comses.net/codebases/26f3d0de-6965-4915-ba81-8ddbcb27b037/releases/1.0.0/`

Acknowledgements

This project was funded by the EPSRC Centre for Doctoral Training in Cybersecurity (EPSRC Grant no EP/S022503/1).

Appendix A: Notation

Notation	Description
$a, \mathcal{A}$	action, set of actions
$b, \|U\|$	number of regular agents
c, C	neighbour, set of neighbours
d, D	service provider, set of service providers
E	set of edges
G	graph
h	feedback
$j, \|M\|$	number of malicious agents
$k, \|D\|$	number of service providers
l	service provider's end-point state
m, M	malicious agent, set of malicious agents
n	service selection rate
$o, \mathcal{O}$	opinion, set of opinions
Q	action-value function
$r, \mathcal{R}$	reward, reward function
$s, \mathcal{S}$	state, set of states
t, T	time step, number of time steps
u, U	regular agent, set of regular agents
v, V	vertex, set of vertexes
W	warm-up time
x	independent variable
y	dependent variable
z	service provider's central state
α	learning rate
ϵ	exploration parameter
ζ	probability of detecting a cyberattack
η	probability of detecting a misinformation campaign
θ	end-point state transition probability $0 \to 1$
κ	number of neighbours in social network
μ	policy for opinion selection
ν	regret
ξ	positive initial value associated with reconnaissance
π	policy for action selection
ρ	rewiring probability
σ	end-point state transition probability $1 \to 0$
τ	satisfaction threshold
υ	central state transition probability $1 \to 0$
Φ	opinion-value function
ψ	central state transition probability $0 \to 1$
ω	direct experience weight

Table 6: Notation.

Appendix B: Sensitivity Analysis

For validating agent-based models, methods such as comparison to real-world data, cross-validation, and sensitivity analysis are typically used (Hunter & Kelleher 2020). However, not all of these methods may be feasible or necessary. In our case, we cannot perform real-data analysis due to the scarcity of empirical data on hybrid threats, specifically on coordinated cyber and misinformation attacks. Our study focuses on the intersection of these two threat types rather than examining them in isolation. As a result, our findings are likely to differ from data on independent cyberattacks or misinformation campaigns since attackers with finite resources would either concentrate them on a single attack or divide them across multiple attacks, affecting each attack's duration, intensity, or frequency. Additionally, cross-validation is not feasible due to the lack of validated models on hybrid threats required for benchmarking.

To validate our model, we used sensitivity analysis. Specifically, we employed regression-based sensitivity analysis (ten Broeke et al. 2016), focusing on key outputs: average service selection rate ($\overline{n}_d$), average service level

($\bar{l}_d$), average positive/negative opinion value ($\overline{o_d^+}$, $\overline{o_d^-}$), average regret ($\overline{\nu}$), cyberattack duration (x_1), misinformation duration (x_3), combined attack duration (x_5), cyberattack start (x_2), misinformation start (x_4), and combined attack start (x_6). Finding a low correlation for $\overline{n}_d$, x_2, x_4, and x_6 (value < 0.20) with all parameters, we focused on the remaining variables. Initial correlation analysis and scatterplots revealed that $\bar{l}_d$ was most influenced by end-point transition probability $0 \rightarrow 1$ (θ), central state transition probability $1 \rightarrow 0$ (υ), and central state transition probability $0 \rightarrow 1$ (ψ), while satisfaction threshold (τ), end-transition probability $1 \rightarrow 0$ (σ), and θ affected both $\overline{o_d^+}$ and $\overline{o_d^-}$, parameters θ, ψ, and υ were most important for $\overline{\nu}$, cyberattack detection probability (ζ) affected x_1 and x_5 and misinformation detection probability (η) influences x_3 and x_5. Table 7, which includes models with R^2 higher than 0.50, details these relationships and Figure 4 visualises them.

Variable	Model	Parameter	Output change	R^2
$\overline{l_d}$	SA 1	θ	0.22 *** (0.014)	0.64
		ψ	0.09 *** (0.014)	
		$\theta : \psi$	0.28 *** (0.022)	
	SA 2	θ	0.44 *** (0.018)	0.63
		υ	-0.21 *** (0.017)	
		$\theta : \upsilon$	-0.11 *** (0.028)	
$\overline{o_d^+}$	SA 3	τ	-0.39 *** (0.080)	0.63
		θ	2.51 *** (0.078)	
		$\tau : \theta$	-3.19 *** (0.136)	
$\overline{o_d^-}$	SA 4	τ	3.00 *** (0.085)	0.53
		σ	1.42 *** (0.089)	
		$\tau : \sigma$	1.88 *** (0.161)	
$\overline{\nu}$	SA 5	θ	0.25 *** (0.009)	0.71
		ψ	0.15 *** (0.009)	
		$\theta : \psi$	0.07 *** (0.014)	
	SA 6	θ	0.11 *** (0.014)	0.60
		υ	-0.27 *** (0.013)	
		$\theta : \upsilon$	-0.30 *** (0.021)	
x_5	SA 7	ζ	-17.41 *** (0.494)	0.51
		η	-20.59 *** (0.470)	
		$\zeta : \eta$	-25.88 *** (0.941)	
x_1		ζ	-18.52 *** (0.455)	0.38
x_3		η	-22.72 *** (0.638)	0.33

Table 7: Sensitivity analysis results. Standard errors in parentheses. *** $p \leq 0.001$; ** $p \leq 0.01$; * $p \leq 0.05$; † $p \leq 0.1$

As expected, higher θ and ψ increase service levels ($\overline{l_d}$) (SA1 on Figure 4). This aligns with the intuitive notion that more availability leads to more successful service experiences. However, a significant negative interaction between υ and θ highlights that central state unavailability has a stronger impact, disrupting service delivery at the end-points (SA2 on Figure 4).

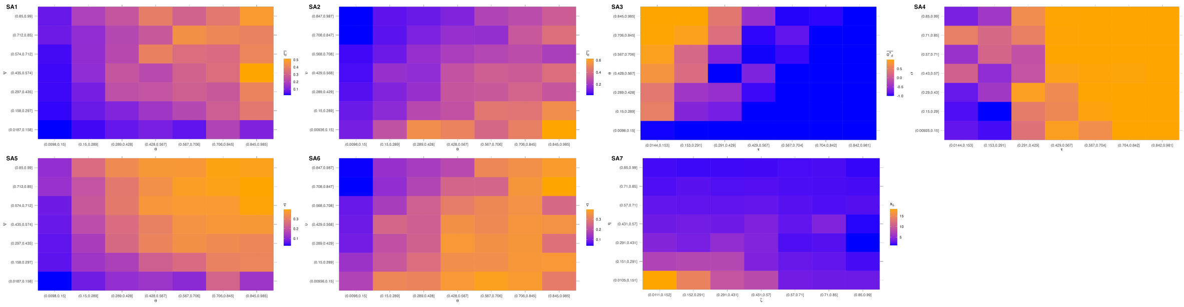

Figure 4: Visualisation of sensitivity analysis results.

A higher τ leads to a decrease in positive opinion value ($\overline{o_d^+}$) while and a higher θ increases $\overline{o_d^+}$. Yet the negative coefficient for $\tau : \theta$ indicates that even when service availability increases if the requirements for satisfaction also increase, people think less positively and more negatively about the service provider (SA3 on Figure 4). A higher τ and σ increase negative opinion value ($\overline{o_d^-}$). Higher values for these parameters would make people believe that the service provider's behaviour is more negative than positive (SA4 on Figure 4).

We also found that increased θ and ψ lead to higher regret ($\overline{\nu}$). This likely arises because higher availability increases the potential for positive rewards during normal operation. Disruption by attacks creates greater contrast, leading to higher regret (SA5 on Figure 4). While θ increases regret, υ lowers regret. Considering the interaction term $\theta : \upsilon$, a lower θ and higher υ would result in lower regret because, similarly to $\theta : \psi$, a lower likelihood of availability means that it is not likely to receive positive rewards with any strategy, minimising the effects of attacks (SA6 on Figure 4).

As expected, higher ζ and η decrease combined attack duration (x_5) (SA7 on Figure 4). Additionally, ζ decreases the duration of a cyberattack (x_1) and η decreases the duration of a misinformation campaign (x_3). These results show that higher detection probabilities decrease the number of time steps attackers can conduct an attack.

References

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., Schneider, J., Tezak, N., Tworek, J., Welinder, P., Weng, L., Yuan, Q., Zaremba, W. & Zhang, L. (2019). Solving Rubik's Cube with a robot hand. arXiv preprint. Available at: `https://arxiv.org/abs/1910.07113`

Azani, E. (2013). The hybrid terrorist organization: Hezbollah as a case study. *Studies in Conflict & Terrorism*, *36*(11), 899–916

Balcaen, P., Bois, C. D. & Buts, C. (2022). A game-theoretic analysis of hybrid threats. *Defence and Peace Economics*, *33*(1), 26–41

Banisch, S., Gaisbauer, F. & Olbrich, E. (2022). Modelling spirals of silence and echo chambers by learning from the feedback of others. *Entropy*, *24*(10)

Banisch, S. & Olbrich, E. (2019). Opinion polarization by learning from social feedback. *The Journal of Mathematical Sociology*, *43*(2), 76–103

Bresniker, K., Gavrilovska, A., Holt, J., Milojicic, D. & Tran, T. (2019). Grand challenge: Applying artificial intelligence and machine learning to cybersecurity. *Computer*, *52*(12), 45–52

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Heigeartaigh, S. O., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R. & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint. Available at: `https://arxiv.org/pdf/1802.07228`

Chowdhary, A., Huang, D., Mahendran, J. S., Romo, D., Deng, Y. & Sabur, A. (2020). Autonomous security analysis and penetration testing. 2020 16th International Conference on Mobility, Sensing and Networking (MSN)

DST Group (2022). CybORG. Available at: `https://github.com/cage-challenge/CybORG`

Elliott, E. O. (1963). Estimates of error rates for codes on burst-noise channels. *Bell System Technical Journal*, *42*(5), 1977–1997

European Commission (2016). Joint framework on countering hybrid threats. Available at: `https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52016JC0018&from=EN`

Giannopoulos, G., Smith, H. & Theocharidou, M. (2020). The landscape of hybrid threats: A conceptual model. Available at: `https://euhybnet.eu/wp-content/uploads/2021/01/Conceptual-Framework-Hybrid-Threats-HCoE-JRC.pdf`

Gilbert, E. N. (1960). Capacity of a burst-noise channel. *Bell System Technical Journal*, *39*(5), 1253–1265

Gunneriusson, H. & Ottis, R. (2013). Cyberspace from the hybrid threat perspective. *Journal of Information Warfare*, *12*(3), 67–77

Hammar, K. & Stadler, R. (2022). Intrusion prevention through optimal stopping. *IEEE Transactions on Network and Service Management*, *19*(3), 2333–2348

Hao, J., Sun, J., Cai, Y., Yu, C. & Huang, D. (2015). Heuristic collective learning for efficient and robust emergence of social norms. Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS) 2015, Richland, SC

Hoffman, F. G. (2007). *The rise of hybrid wars*. Arlington: Potomac Institute for Policy Studies

Hoffman, F. G. (2010). 'Hybrid threats': Neither omnipotent nor unbeatable. *Orbis*, *54*(3), 441–455

Huang, L. & Zhu, Q. (2020). A dynamic games approach to proactive defense strategies against Advanced Persistent Threats in cyber-physical systems. *Computers and Security*, *89*, 1–16

Hunter, E. & Kelleher, J. (2020). A framework for validating and testing agent-based models: A case study from infectious diseases modelling. 4th. Annual European Simulation and Modelling Conference

Jacobs, A. & Samaan, J.-L. (2015). Player at the sidelines: NATO and the fight against ISIL. NATO's Response to Hybrid Threats, NATO Defense College, Rome

Keith, A. & Ahner, D. (2021). Counterfactual regret minimization for integrated cyber and air defense resource allocation. *European Journal of Operational Research*, *292*(1), 95–107

Krotofil, M., Cárdenas, A., Larsen, J. & Gollmann, D. (2014). Vulnerabilities of cyber-physical systems to stale data - Determining the optimal time to launch attacks. *International Journal of Critical Infrastructure Protection*, *7*, 213–232

Kumar, K. K. & Geethakumari, G. (2013). Information diffusion model for spread of misinformation in online social networks. 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), New York

Kurt, M. N., Ogundijo, O., Li, C. & Wang, X. (2019). Online cyber-attack detection in smart grid: A reinforcement learning approach. *IEEE Transactions on Smart Grid*, *10*, 5174–5185

Linkov, I., Baiardi, F., Florin, M.-V., Greer, S., Lambert, J. H., Pollock, M., Rickli, J.-M., Roslycky, L., Seager, T., Thorisson, H. & Trump, B. D. (2019). Applying resilience to hybrid threats. *IEEE Security and Privacy*, *17*(5), 78–83

Microsoft (2021). CyberBattleSim. Available at: `https://github.com/microsoft/CyberBattleSim`

Mueller, D. C. (1996). Public choice in perspective. In D. C. Mueller (Ed.), *Perspectives on Public Choice*, (pp. 1–18). Cambridge: Cambridge University Press

Mukherjee, P., Sen, S. & Airiau, S. (2008). Norm emergence under constrained interactions in diverse societies. Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, Richland, SC

Myerson, R. B. (1991). *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press

NATO Strategic Communications (2020). Hybrid threats: A strategic communications perspective. Available at: `https://stratcomcoe.org/hybrid-threats-strategic-communications-perspective`

Oh, S. H., Jeong, M. K., Kim, H. C. & Park, J. (2023). Applying reinforcement learning for enhanced cybersecurity against adversarial simulation. *Sensors*, *23*(6)

Perra, N. & Rocha, L. E. C. (2019). Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific Reports*, *9*(1), 7261

Podobnik, B., Horvatic, D., Lipic, T., Perc, M., Buldú, J. M. & Stanley, H. E. (2015). The cost of attack in competing networks. *Journal of The Royal Society Interface*, *12*(112), 1–12

Poole, D. L. & Mackworth, A. K. (2017). *Artificial Intelligence: Foundations of Computational agents*. Cambridge: Cambridge University Press

Sayama, H. (2015). *Introduction to the Modeling and Analysis of Complex Systems*. Geneseo, NY: Open SUNY Textbooks

Sazonov, V., Mölder, H., Müür, K., Pruulmann-Vengerfeldt, P., Kopõtin, I., Ermus, A., Salum, K., Šlabovitš, A., Veebel, V. & Värk, R. (2016). Russian information campaign against the Ukrainian state and defence forces. NATO Strategic Communications Centre of Excellence, Estonian National Defence College

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K. & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144

Sinha, A., Nguyen, T. H., Kar, D., Brown, M., Tambe, M. & Jiang, A. X. (2015). From physical security to cybersecurity. *Journal of Cybersecurity*, *1*(1), 19–35

Smith, M. J. & Price, G. R. (1973). The logic of animal conflict. *Nature*, *246*(5427), 15–18

Sutton, R. S. & Barto, A. G. (2020). *Reinforcement Learning: An Introduction*. Cambridge, MA: The MIT Press

ten Broeke, G., van Voorn, G. & Ligtenberg, A. (2016). Which sensitivity analysis method should I use for my agent-based model? *Journal of Artificial Societies and Social Simulation*, *19*(1), 5

Treverton, G. F. (2018). The intelligence challenges of hybrid threats: Focus on cyber and virtual realm. Tech. rep., Swedish Defence University, Bromma

Treverton, G. F., Thvedt, A., Chen, A. R., Lee, K. & McCue, M. (2018). Addressing hybrid threats. Swedish Defence University

UK Government (2015). National security strategy and strategic defence and security review 2015: A secure and prosperous United Kingdom. HM Government

Wang, F.-Y. (2010). The emergence of intelligent enterprises: From CPS to CPSS. *IEEE Intelligent Systems*, *25*(4), 85–88

Wang, Y., Yang, W., Ma, F., Xu, J., Zhong, B., Deng, Q. & Gao, J. (2020). Weak supervision for fake news detection via reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 516–523

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440–442

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*(3–4), 330–336

Yu, C., Tan, G., Lv, H., Wang, Z., Meng, J., Hao, J. & Ren, F. (2016). Modelling adaptive learning behaviours for consensus formation in human societies. *Scientific Reports*, *6*(1), 27626

Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N. J., Xie, X. & Li, Z. (2018). DRN: A deep reinforcement learning framework for news recommendation. Proceedings of the 2018 World Wide Web Conference, Republic and Canton of Geneva, CHE

Zhou, S., Liu, J., Hou, D., Zhong, X. & Zhang, Y. (2021). Autonomous penetration testing based on improved deep Q-network. *Applied Sciences*, *11*(19)