Securing Autonomous Vehicle Networks: Strategic Defence Against Communication Attacks

Leandro Parada and Panagiotis Angeloudis
Centre for Transport Studies
Department of Civil and Environmental Engineering
Imperial College London, UK
{lap20, p.angeloudis}@imperial.ac.uk

***Abstract*—Adversarial communication poses a significant threat to cooperative autonomous driving, with subtle attacks potentially resulting in catastrophic consequences. This paper addresses the vulnerabilities of Multi-Agent Reinforcement Learning (MARL) systems in autonomous driving to communication attacks and introduces TrustCommNet—a novel defence framework. TrustCommNet secures MARL communication by integrating a Theory of Mind predictor with a trust-based validation mechanism, advancing beyond traditional detection and prediction methods to actively mitigate adversarial threats. Our rigorous evaluations demonstrate how communication perturbations can severely compromise the performance and safety of autonomous driving systems. TrustCommNet exhibits a robust defence, achieving zero-shot resilience in most scenarios without prior exposure to the attack, and suffers 50% less average disruption, quickly restoring cooperative performance. Comprehensive experiments confirm that TrustCommNet scales effectively against numerous attackers, maintaining near-perfect success rates in fast-paced urban environments.**

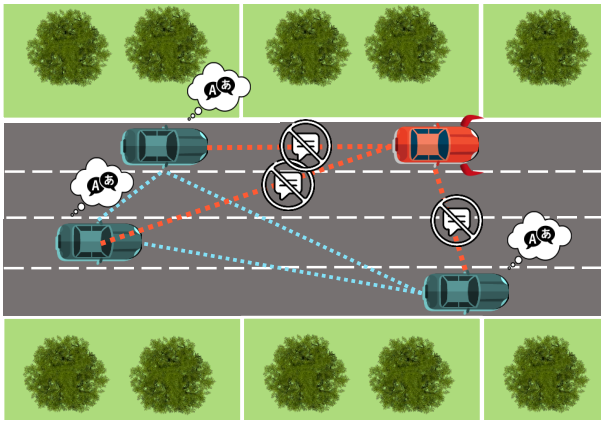


Fig. 1: Conceptual overview of the proposed defence mechanism to maintain network integrity by detecting adversaries and reconstructing their messages.

I. Introduction

Adversarial communication poses a significant threat to multi-agent systems, undermining their collaborative efficiency and compromising security. This emerging challenge, crucial in the context of cooperative autonomous driving, necessitates a deeper understanding of how deceptive information impacts multi-agent operations and the development of effective countermeasures. Unlike traditional adversarial attacks that primarily focus on manipulating a single ego vehicle's perception or decision-making process [36, 13, 23], communication attacks exploit the interdependence of V2X communication protocols [31].

Adversarial attacks on V2X communications are a relatively new and under-researched field, primarily addressed through imitation learning. Previous studies have shown that even simple perturbations can lead to potent attacks, and adversarial training is effective when the attack model is known [31]. To defend against V2V attackers, methods have been developed that involve sampling a subset of messages and verifying consensus between object detections [14]. In a recent study, a mechanism was developed to generate perturbations in LiDAR-based V2X systems, and training on these perturbed scenes significantly enhanced performance [34]. These methods involve pre-generating adversarial images or point clouds to then train using imitation learning, which lacks adaptability. In contrast, framing the problem as a communicative Multi-Agent Reinforcement Learning (MARL) problem allows for training adaptive attackers, leading to the development of more robust defence mechanisms.

Recent advancements in the field of communication-based MARL have demonstrated the immense benefits of these methods, particularly in enhancing the coordination and overall performance of multiple agents [37]. This burgeoning area of research focuses on developing and implementing strategies that range from broadcasting simple messages [5, 29] to targeted communications for specific groups [3, 20, 16]. Nonetheless, the increasing prevalence of adversarial threats in MARL calls for new solutions that ensure secure communication, striking a balance between efficient collaboration and strong defences.

Despite the limited research on adversarial communication in MARL, key studies have begun to shed light on the complexity of defence mechanisms and adversarial strategies. For instance, Blumenkamp et al. [1] observed that agents can develop adversarial communication strategies driven by competitive incentives rather than adversarial design. Xue et

al. [35] developed deliberative adversarial communication methods, while Mitchell et al. [17] and Tu et al. [31] explored message detection methods and adversarial attacks on feature maps, highlighting significant vulnerabilities. Despite advancements in understanding adversarial communication and detection, practical defence strategies are limited. Our research aims to bridge this gap by introducing a modular defence architecture that extends beyond traditional detection and prediction to mitigate adversarial threats.

This paper presents TrustCommNet, an end-to-end learnable architecture that enhances autonomous driving security by ensuring communication integrity within autonomous vehicle networks. TrustCommNet proactively defends against adversarial threats through real-time checks and adaptive responses. Its robustness comes from key components: a system to validate communication integrity and filter malicious content, a message scheduling system for effective timing and targeting, and a message reconstruction process to correct distorted messages.

To test our proposed defence mechanism, we craft an adaptive attacker model based on Projected Gradient Descent (PGD) [10] to execute ℓ_∞-bounded perturbations on communication messages. We use this model to demonstrate how perturbations in communication can significantly impact team performance, which leads to poor success rates for state-of-the-art algorithms [20, 28]. We then test the resilience of our defence framework in three urban driving scenarios: Highway, Intersection and Lane merging. The results demonstrate our framework's capability to detect and counteract sophisticated attacks effectively, quickly recovering from disruptions and restoring prior levels of cooperative performance. Through various experiments, we show that our framework achieves near-perfect zero-shot performance in challenging environments and can scale to accommodate multiple attackers.

II. Related Work

A. Multi-Agent Reinforcement Learning with Communication.

Recent advancements in MARL emphasise the critical role of efficient communication strategies among agents to achieve cooperative objectives while minimising bandwidth use. A comprehensive survey can be found in [37]. Early works such as DIAL [5], RIAL [5] and CommNet [29] introduced the concept of learning to communicate in MARL with fully connected structures, ideal for environments requiring collective decision-making. A few years later, ATOC [8] and IC3Net [28] introduced gate mechanisms, enabling selective communication based on agents' local context, enhancing efficiency in scenarios where continuous information exchange is unnecessary. TarMAC [3] leverages an attention mechanism for targeted communication, ensuring relevance and efficiency in message exchange. SchedNet [32] and GA-Comm [16] utilise communication graphs, with SchedNet focusing on agent importance to manage bandwidth and GA-Comm incorporating attention mechanisms for refined communication processes. MAGIC [20], also adopting a communication graph, adapts dynamically to changing agent states and environmental conditions.

B. Adversarial Attacks on Deep Learning and Communication.

Adversarial attacks on deep learning models, initially recognised in image classification [30, 7, 19], have since extended into areas like multi-agent reinforcement learning (MARL) [4, 9] and communication [31, 35, 1]. These attacks are categorised into white-box [22], where attackers have full knowledge of the network, and black-box [2, 31], executed without detailed network information. Despite extensive exploration, the specific targeting of multi-agent communication within MARL remains under-investigated.

Recent studies have made significant strides in understanding adversarial communication within MARL. Mitchell et al. [17] introduced a method to evaluate message reliability using a modified Graph Neural Network (GNN) layer in attention-based MARL systems. Blumenkamp et al. [1] discovered that agents naturally develop adversarial communication strategies in competitive-cooperative environments. In contrast, Xue et al. [35] studied deliberate adversarial communication and proposed a defence mechanism conceptualising it as a two-player zero-sum game. The work of Tu et al. [31] highlights vulnerabilities in distributed deep learning, showing adversarial training's effectiveness when the threat model is known. Fung et al. [6] proposed a decentralised trust mechanism for MARL, focused on binary consensus tasks.

These studies highlight the complexity of adversarial strategies in MARL and the need for robust defences. The majority of the literature focuses on the generation or detection of adversarial communication, with few works on defence mechanisms. The most notable work is probably the one by Xue et al. [35]. In contrast, our framework not only detects adversarial messages but also validates, schedules, and reconstructs communications to maintain integrity and cooperative performance.

C. Adversarial Attacks on V2X communications.

Adversarial attacks on V2X communications represent a relatively new field with limited research. Notable work includes the study by Tu et al. [31], which investigated adversarial transfer attacks on collaborative perception based on LiDAR measurements. Their findings demonstrated that even simple perturbations could result in potent attacks and that adversarial training is an effective defence mechanism when the attack model is known. Li et al. [14] proposed an intelligent sampling method designed to aid CAVs in achieving consensus in the presence of V2V attackers. This method involves sampling a subset of messages at each

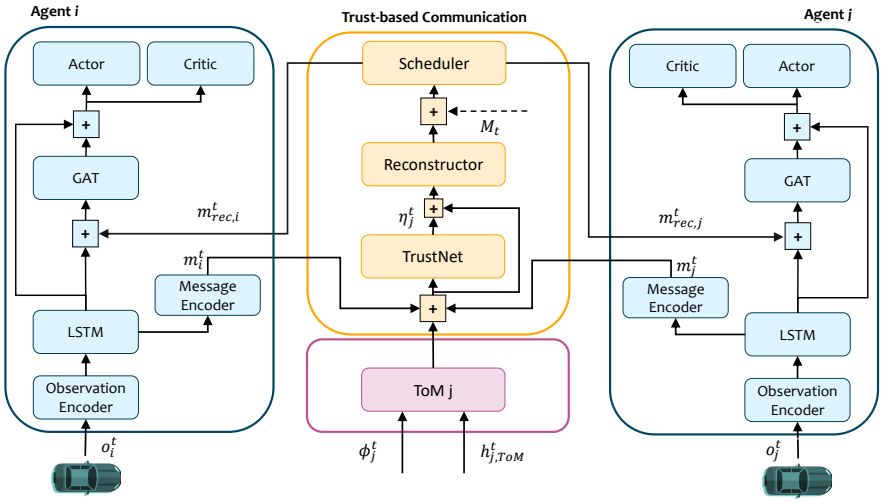


Fig. 2: This schematic presents TrustCommNet, our proposed defence architecture for autonomous vehicle networks, highlighting its key components: a Theory of Mind predictor for agent interactions, a message validation and reconstruction mechanism, and a scheduler for efficient bandwidth use.

step and verifying consensus between two object detections. Similarly, Xiang et al. [34] developed a mechanism capable of generating perturbations in LiDAR-based V2X perception systems and demonstrated that training on these perturbed scenes can significantly enhance system performance. These studies underscore the importance of developing robust defences and innovative methods to mitigate the impact of adversarial attacks on V2X communications.

D. Machine Theory of Mind.

The concept of Machine Theory of Mind (MToM) has been explored in multi-agent systems, focusing on interpreting behaviours and enhancing cooperation. Rabinowitz et al. [24] introduced ToMNet, a meta-learning method to predict agent behaviours based on local observations. Shu et al. [26] proposed a hierarchical approach for a manager agent to coordinate worker agents by inferring their preferences. Shu et al. [27] presented AGENT, a benchmark to assess machine agents' understanding of intuitive psychology. Sclar et al. [25] created an environment requiring agents to model others' mental states for high rewards. Wang et al. [33] proposed ToM2C, a method to decide when and with whom to communicate in multi-agent environments by predicting others' observations and goals.

Our approach uniquely predicts other agents' communication states, unlike traditional methods focused on observations or goals. Using only positional data and agent characteristics for message prediction, it addresses scenarios where agents lack access to others' observations or actions, showcasing a novel ToM application in multi-agent analysis.

III. Problem Formulation

A. Communication-Based MARL

We explore communication-based MARL as an extension of a Decentralised Partially Observable Markov Decision Process (Dec-POMDP) [21]. This framework is formally represented by the tuple $\langle N, S, A, O, \mathcal{T}, R, \gamma, \mathcal{M}, \mathcal{C} \rangle$. Within this framework, N denotes a set of agents $N = \{1, 2, \ldots, n\}$, S represents the global state space, and A encompasses the joint action space, with each agent i having individual actions $a_i \in A$ and observations $o_i \in O$. The system's dynamics are governed by a transition function $\mathcal{T} : S \times A \to S$, mapping the current state and joint actions to the next state, and a reward function $R : S \times A \times S \to \mathbb{R}$, which assigns a scalar reward to each state-action pair.

A key feature of communication-based MARL is the incorporation the set of possible messages $\mathcal{M} = \{m_1, m_2, \ldots, m_k\}$, and the communication protocol or aggregation function $\mathcal{C} : \mathcal{M}^N \to \mathcal{M}$, which consolidates individual agents' messages into a unified message. Each agent has two policies, an action policy $\pi_i(a_i|o_i, m_{rec,i})$, and a message policy $\xi_i(m_i|o_i, m_{rec,i})$, where $m_{rec,i}$ is the message received by agent i, consolidated by the communication protocol. The goal of each agent is to maximise the expected cumulative discounted reward, as defined by the objective function $\max_{\pi_1,\pi_2,\ldots,\pi_N} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r^t(s^t, a_1^t, a_2^t, \ldots, a_N^t)\right]$, where γ is the discount factor and T the time horizon.

B. Adversarial MARL

Building on the Byzantine Generals Problem, which explores consensus difficulties in the presence of unreliable agents [11], the Adversarial MARL approach addresses the

uncertainty regarding the number and strategies of adversarial agents. We consider $N_{adv} \subseteq N$ adversarial agents equipped with the same observational capabilities as standard agents but are also capable of crafting malicious messages m_{adv} to mislead others, thereby disrupting decision-making and compromising collective goals. Their behaviour is guided by an adversarial message policy $\xi_{adv}(m_{adv}|\tau, m)$, where τ is the history of observations and actions and m is the unaltered message produced by the agent through the message policy. Thus, the adversarial agents aim to introduce carefully crafted message alterations to damage the overall team performance.

IV. Methodology

In this section, we present out method for countering adversarial attacks on communication in cooperative MARL. Our framework includes an adversarial algorithm using ℓ_∞-bounded perturbations using PGD with the objective of reducing the aggregated reward. For defence, we employ TrustCommNet, a trust-based validation network to ensure message integrity, supported by a ToM-like predictor. The framework's architecture is illustrated in Figure 2.

A. Dynamic ℓ_∞-bounded Attack on Communication

Designing a strategic DRL attacker mainly involves addressing two challenges: determining "how" and "when" to launch attacks. Our "how-to-attack" strategy introduces subtle message perturbations to create a disparity between the network's predicted and actual outputs, systematically deviating agents' policies from their optimal paths. Utilising the differentiability inherent in the system, this method efficiently computes necessary gradients without needing full environmental knowledge.

We use PGD to optimise the adversarial message m_{adv} and target the team loss within the MARL context. Employing PGD enables precise manipulation of the agent's messages to maximise the policy loss while operating within the ℓ_∞-bound, thus striking a crucial balance between the efficacy of the attack and undetectability. The PGD update for the adversarial message is articulated as follows:

$$m^{(t+1)} = Proj_{m+\epsilon}\left(m^t + \alpha \cdot sgn\left(\nabla_{m^t} L(\theta, m^t, h^t)\right)\right) \tag{1}$$

where $m^{(t+1)}$ is the updated adversarial message, with sgn denoting the sign function, and h^t as the current hidden state. The step size is given by α, and $Proj_{m+\epsilon}$ maintains the message within the allowable perturbation bounds, ϵ. We utilise a cross-entropy loss function that maximises the probability of the worst possible actions. This approach guides adversarial agents to craft messages m_{adv} that effectively degrade the performance of non-adversarial agents by increasing the likelihood of sub-optimal decisions.

We now shift focus to the "when-to-attack" problem, which introduces an additional layer of complexity to the task of identifying potential attackers within a system. Our methodology for timing attacks hinges on a specifically designed preference function for the attacker. We adopt a policy output disparity function, similar to [15]. This function evaluates the disparity between the maximum and minimum outputs of an agent's policy at a given step:

$$c = \max \pi_i(o_i^t, m_{rec,i}^t) - \min \pi_i(o_i^t, m_{rec,i}^t) \tag{2}$$

The rationale behind employing this function is to evaluate the decisiveness of a DRL agent towards a particular action; a larger difference indicates a strong preference for a specific action, justifying an attack. Conversely, a minimal difference suggests an absence of clear preference, advising against an attack on this time step. To operationalise this strategy, we introduce a threshold β, serving as the critical value for $c(o_i^t)$ above which an attack is deemed worthwhile.

We acknowledge there are more complex attacking strategies in the literature, like state luring mechanisms [15, 18]; however, our focus in this paper is on the defence mechanism. Thus, we just want to create a simple yet effective attack on multi-agent communication.

B. Trust Verification Defence Mechanism

Building on the concept of Machine Theory of Mind (ToM) [24], our model enables agents to predict the communication messages of others. While similar to the ToM2C approach [33], which infers goals and observations from pose and encoded observation, our method predicts encoded communication messages of other agents based solely on their pose ϕ_j.

First, as illustrated in figure 2, the imagined message for an external agent j from the point of view of an agent i is $f_{ToM,j,i}(m_{j,i,ToM}^t|\phi_j^t, h_{j,i,ToM}^t)$. This model considers as input agent's j pose ϕ_j^t, and a hidden state $h_{j,i,ToM}^t$ that represents the history of past behaviour. From this point onwards, we drop the index of agent i for simplicity. It should be noted that figure 2 only displays the modelling network of agent j from the perspective of agent i, $f_{ToM,j,i}$. However, each agent has an independent network to model the communications of others. Since all agents' communication messages have the same form, they can share the weights of the network $f_{ToM,j}$. In practice, we use an LSTM followed by a fully connected layer to represent $f_{ToM,j}$.

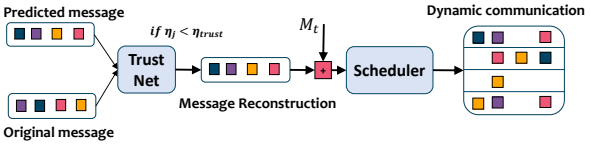


Fig. 3: Trust-based mechanism for detecting, reconstructing and scheduling messages in the communication network.

Second, we introduce and train a network layer, which we name TrustNet, designed to assess the reliability of each agent's communication. TrustNet operates by comparing the predicted communication message with the actual message sent by each agent. Specifically, for each agent j at time t, TrustNet evaluates $f_{Trust}(\eta_j^t | m_j^t, m_{j,ToM}^t)$, where η_j^t acts as a probabilistic trust indicator. This indicator approximates the agent's reliability by quantifying the discrepancy between its expected message $m_{j,ToM}^t$ and the actual communication message m_j^t. We assume the homogeneity of all agents' observations and messages. Thus, we assign the same network parameters for f_{ToM} and f_{Trust} across agents.

To safeguard the integrity of multi-agent communications, our framework implements a dynamic trust-based protocol. In instances where an agent's trust score descends below a trust threshold $\eta_j^t < \eta_{trust}$, the system activates protective protocols and the agent's communication message will then be blocked and a new message can be reconstructed based on the predicted message $m_{j,\text{ToM}}^t$ and agent's i current observation o_i:

$$m_{j,\text{new}}^t = \begin{cases} f_{Rec}(m_{j,\text{ToM}}^t, o_i) & \text{if } \eta_j^t < \eta_{\text{trust}} \\ m_j^t & otherwise, \end{cases} \tag{3}$$

Following message reconstruction, a communication scheduler processes the updated messages from all agents, generating an adjacency matrix G^t that defines the communication scheme using a hard attention mechanism. To maintain differentiability during training, a Gumbel-Softmax layer was employed. The matrix G^t indicates active communication with $G_{ij}^t = 1$ for agent i to agent j at time t and $G_{ij}^t = 0$ otherwise. A simplified schematic overview of the trust-based message detection, reconstruction, and scheduling process is shown in figure 3.

Agents then receive incoming aggregated messages $m_{rec,i}^t$ and integrate them with their own encoded observation using a Graph Attention Network (GAT). This is then utilised to predict actor and critic outputs as depicted in figure 2. This structured approach allows agents to dynamically refine their communication strategies, enhancing message efficiency and the robustness of the network against potential threats.

C. Overall Training Objective

To enable agents to develop accurate predictions of other agents' communications, we use the Kullback-Leibler (KL) divergence between the predicted and actual communication probability:

$$\mathcal{L}_{ToM} = \frac{1}{N} \sum_{j=1}^{N} \sum_{i \neq j}^{N-1} D_{KL}(p(m_j^t) || q(m_{j,ToM}^t | \phi_j^t, h_{j,ToM}^t)) \tag{4}$$

where N is the number of agents, $p(m_j^t)$ is the actual communication probability distribution for the j-th message sample, and $q(m_{j,ToM}^t | \phi_j^t, h_{j,ToM}^t)$ is the corresponding predicted distribution by the ToM layer. The KL divergence-based loss is crucial for accurately measuring differences between the model's predicted probabilities and actual communication behaviours, ensuring effective learning of communication dynamics in the multi-agent environment.

Additionally, the training process employs an Actor-Critic loss ($\mathcal{L}_{AC}$) to optimise the agents' policies towards maximising expected discounted rewards. This loss integrates a policy gradient component, optimising the agents' policy networks, and a value component, refining the value function estimates:

$$\begin{aligned} \mathcal{L}_{AC} = & \mathbb{E}_{(\tau, m^{adv})} \left[\sum_{t=0}^{T} A^t \cdot \nabla_\theta \log \pi_i(a_i^t | o_i^t, m_i^t) \right] \\ & + \lambda \sum_{t=0}^{T} \nabla_\rho (R_i^t - V_\rho(o_i^t))^2 \end{aligned} \tag{5}$$

where $A^t = (R_i^t - V_\rho(o_i^t))$ is the advantage function and V_ρ is the value function with set of parameters ρ. In practice, we let the policy and the value function share the same parameters of the neural network except for their respective output layers (heads). The parameter λ balances the two components, facilitating effective policy learning and decision-making. Therefore, the overall training loss of the model is $L_{tot} = L_{ToM} + L_{AC}$.

We use the Centralised Training Decentralised Execution (CTDE) paradigm. During training, agents utilise observations and communication from others, but during inference, they rely on local observations and the scheduler-determined adjacency matrix G^t. This framework assumes agents know each other's locations to predict communication patterns without accessing others' observations or messages.

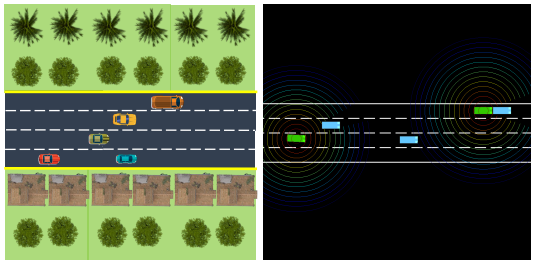
V. Experiments

A. Environment Description

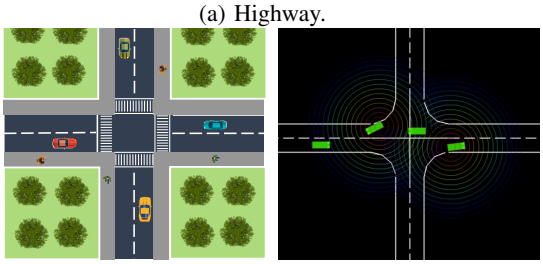
We use a modified version of Highway-env [12], adapted for multi-agent urban navigation where agents avoid collisions and reach their destination to maximise rewards. Our approach introduces randomness in vehicles' speeds, locations, and destinations to enhance adaptability and limits agent communication to 60 meters to maintain realism, aligning with DSRC protocols.

The elements of the gym environment can be summarised as follows:

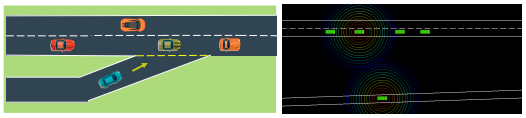
- Agents: CAVs with limited visibility, capable of exchanging messages through a vehicle-to-vehicle communication network.
- Observation: A bespoke LiDAR-like method using ray-casting to segment the 360° field around the agent, detecting other vehicles within an 80-meter range.
- Action: Discrete meta-actions including acceleration (minor and major), deceleration (minor and major), speed maintenance, and lane changes (left and right).



(a) Highway.



(b) Intersection.



(c) Lane merging.

Fig. 4: Scenarios used for the experiments.

- Reward: A cooperative game where rewards are based on speed, collision avoidance, arrival, and bonuses for cooperative lane merging.

B. Result Comparison

In this section, we present the experimental results from testing the adversarial defence capabilities of various state-of-the-art communicative MARL algorithms, including our proposed TrustCommNet. The baseline algorithms evaluated are MAGIC [20] and IC3Net [28], applied to our modified version of highway-env.

Our primary evaluation metric is the average success rate (%), or win rate, across all environments and scenarios, indicating the proportion of episodes successfully completed. Additionally, we assess performance using the average team return. To ensure the robustness of training results, we employ 5 independent seeds to generate learning curves, and we validate our experiments across 100 independent test episodes.

The procedure for adversarial training of the baselines and TrustCommNet is described as follows. Initially, all models were pretrained to achieve clear convergence, attaining an average test success rate of over 99%. After this pretraining, the models were exposed to the attacker for the first time. Subsequently, all models were retrained to assess their inherent resilience and ability to counteract the disruptive effects of the malicious messages. During this phase of the experiment, adversarial messages were incorporated at the 4000 timestep mark.

Figure 5 presents the result comparison of performing adversarial training for 2×10^5 timesteps. Moreover, the zero-shot performance over 100 test episodes of the different algorithms is presented in table I. Our method requires minimal training and, in some cases (intersection and highway), demonstrates zero-shot safety against adversarial threats without previous exposure to the communication attack. In contrast, both MAGIC and IC3Net are severely affected by adversarial communication, displaying high levels of disruption (low success rate) and long recovery tails. IC3Net is not able to recover in two out of three scenarios in 2×10^5, while MAGIC is able to recover performance in all three scenarios but takes several timesteps to reach previous performance levels.

Figure 6 shows TrustCommNet in action in the Intersection scenario with a single adversarial agent (agent 0). The figure shows the communication graph and normalised average trust probability for each agent (heatmap). Initially, all agents share messages to locate the prey (frames (a) and (b)). As the episode progresses, agent 0's malicious behaviour is detected, leading to an increase in its average trust score (heatmap) and subsequent blocking by other agents (frames (c) and (d)). By reconstructing the original message, agents can navigate the intersection safely.

TABLE I: Zero-shot performance of different algorithms when exposed to communication attacks.

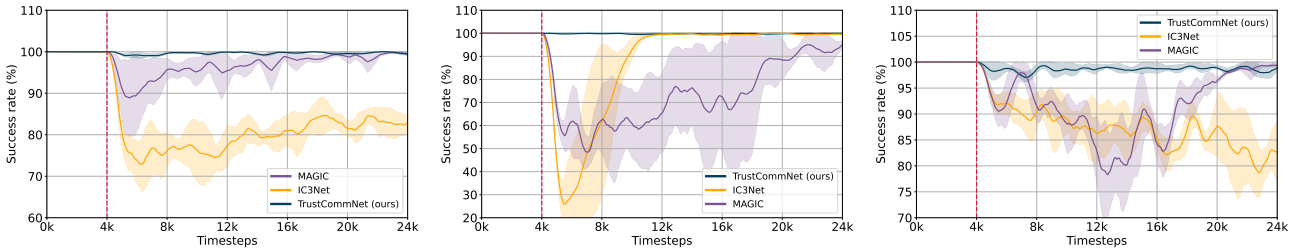
Method	Highway	Intersection	Lane-Merging
IC3Net	75.5 ± 8.5	34.0 ± 12.45	32.5 ± 9.5
MAGIC	90.83 ± 6.4	52.04 ± 25.3	84.0 ± 12.8
TrustCommNet	99.3 ± 1.2	99.0 ± 1.3	97.6 ± 2.1

C. Scaling to Multiple Attackers

We evaluated the performance of our proposed defence mechanism against varying numbers of attackers in an Intersection environment with 8 agents, using the zero-shot success rate (%) as the metric. The results, presented in Table II, show the rapid performance decline of the IC3Net benchmark as the number of attackers increases, while MAGIC exhibits stable but poor performance. In contrast, TrustCommNet maintains robust performance across all attacker counts, demonstrating its potential to scale effectively in larger networks with multiple attackers. These findings are also depicted graphically in Figure II.

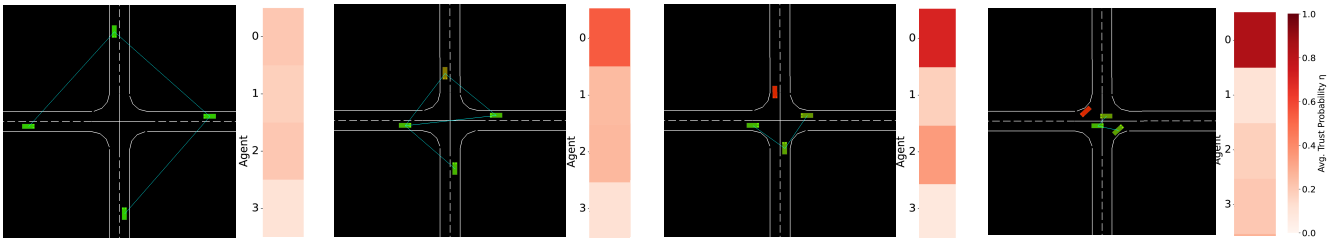
VI. Conclusion

This study presents TrustCommNet, a novel framework developed to enhance the robustness of autonomous driving systems against adversarial communication. We model the adversarial communication problem as a Multi-Agent Reinforcement Learning problem and devise a defence mechanism that is able to counteract the attacks. Our



(a) Highway (b) Intersection (c) Lane Merging

Fig. 5: Performance comparison of adversarial training between TrustCommNet and baseline models. Adversarial communication and training commenced at the 4000-step mark, indicated by the dashed line.



(a) (b) (c) (d)

Fig. 6: Sequence of an episode of the Intersection scenario featuring a single attacker agent (agent 0). The figures illustrate part of the defence strategy, which involves blocking the communication of the corrupted agent.

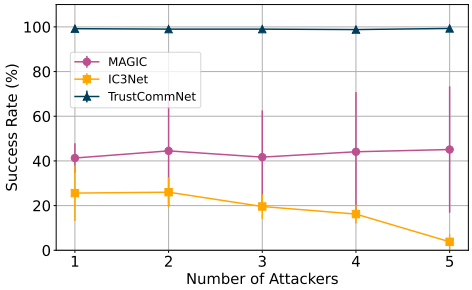


Fig. 7: Scalability of TrustCommNet and Baselines with varying numbers of attackers.

TABLE II: Scalability of zero-shot algorithm defence across different numbers of attackers.

N° Attackers	MAGIC	IC3Net	TrustCommNet
1	41.3 ± 6.6	25.6 ± 12.45	99.2 ± 1.4
2	44.5 ± 24.7	26.0 ± 6.7	99.0 ± 1.3
3	41.7 ± 20.9	19.6 ± 5.5	99.0 ± 1.3
4	44.1 ± 26.7	16.2 ± 4.2	98.8 ± 1.3
5	45.1 ± 28.29	3.8 ± 3.3	99.3 ± 1.2

comprehensive evaluation against ℓ_∞-bounded adversarial attack demonstrates how message alterations can significantly diminish team performance in state-of-the-art MARL algorithms. Although baseline frameworks provide certain defences by adversarial training, this process often disrupts network cooperation, leading to lower success rates. In contrast, TrustCommNet shows near-perfect zero-shot performance in all driving scenarios, surpassing retraining conventional communication MARL algorithms. Future efforts should concentrate on scaling up and refining communication protocols to reinforce security in complex, agent-dense environments, and will also assess different types of attacks.

References

[1] Jan Blumenkamp and Amanda Prorok. The Emergence of Adversarial Communication in Multi-Agent Reinforcement Learning, November 2020. URL http://arxiv.org/abs/2008.02616. arXiv:2008.02616 [cs].

[2] Thomas Brunner, Frederik Diehl, Michael Truong Le, and Alois Knoll. Guessing Smart: Biased Sampling for Efficient Black-Box Adversarial Attacks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4957–4965, October 2019. doi: 10.1109/ICCV.2019.00506. URL http://arxiv.org/abs/1812.09803. arXiv:1812.09803 [cs, stat].

[3] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Michael Rabbat, and Joelle Pineau. TarMAC: Targeted Multi-Agent Communication, February 2020. URL http://arxiv.org/abs/1810.11187. arXiv:1810.11187 [cs, stat].

[4] Martin Figura, Krishna Chaitanya Kosaraju, and Vijay Gupta. Adversarial attacks in consensus-based multi-agent reinforcement learning, March 2021. URL http://arxiv.org/abs/2103.06967. arXiv:2103.06967 [cs, eess, stat].

[5] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to Communicate with Deep Multi-Agent Reinforcement Learning, May 2016. URL http://arxiv.org/abs/1605.06676. arXiv:1605.06676 [cs].

[6] Ho Long Fung, Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. Trust-based Consensus in Multi-Agent Reinforcement Learning Systems, May 2022. URL http://arxiv.org/abs/2205.12880. arXiv:2205.12880 [cs].

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples, March 2015. URL http://arxiv.org/abs/1412.6572. arXiv:1412.6572 [cs, stat].

[8] Jiechuan Jiang and Zongqing Lu. Learning Attentional Communication for Multi-Agent Cooperation, November 2018. URL http://arxiv.org/abs/1805.07733. arXiv:1805.07733 [cs].

[9] Kiarash Kazari, Ezzeldin Shereen, and Gyorgy Dan. Decentralized Anomaly Detection in Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 162–170, Macau, SAR China, August 2023. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-03-4. doi: 10.24963/ijcai.2023/19. URL https://www.ijcai.org/proceedings/2023/19.

[10] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Machine Learning at Scale, February 2017. URL http://arxiv.org/abs/1611.01236. arXiv:1611.01236 [cs, stat].

[11] Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. In *Concurrency: the Works of Leslie Lamport*, pages 203–226. Association for Computing Machinery, New York, NY, USA, October 2019. ISBN 978-1-4503-7270-1. URL https://doi.org/10.1145/3335772.3335936.

[12] Edouard Leurent. An environment for autonomous driving decision-making, 2018.

[13] Simin Li, Jun Guo, Jingqiao Xiu, Pu Feng, Xin Yu, Aishan Liu, Wenjun Wu, and Xianglong Liu. Attacking Cooperative Multi-Agent Reinforcement Learning by Adversarial Minority Influence, June 2023. URL http://arxiv.org/abs/2302.03322. arXiv:2302.03322 [cs].

[14] Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, and Chen Feng. Among Us: Adversarially Robust Collaborative Perception by Consensus. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 186–195, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.00024. URL https://ieeexplore.ieee.org/document/10377144/.

[15] Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of Adversarial Attack on Deep Reinforcement Learning Agents, November 2019. URL http://arxiv.org/abs/1703.06748. arXiv:1703.06748 [cs, stat].

[16] Yong Liu, Weixun Wang, Yujing Hu, Jianye Hao, Xingguo Chen, and Yang Gao. Multi-Agent Game Abstraction via Graph Attention Neural Network, November 2019. URL http://arxiv.org/abs/1911.10715. arXiv:1911.10715 [cs].

[17] Rupert Mitchell, Jan Blumenkamp, and Amanda Prorok. Gaussian Process Based Message Filtering for Robust Multi-Agent Cooperation in the Presence of Adversarial Communication, December 2020. URL http://arxiv.org/abs/2012.00508. arXiv:2012.00508 [cs].

[18] Kanghua Mo, Weixuan Tang, Jin Li, and Xu Yuan. Attacking Deep Reinforcement Learning With Decoupled Adversarial Policy. *IEEE Transactions on Dependable and Secure Computing*, 20(1):758–768, January 2023. ISSN 1941-0018. doi: 10.1109/TDSC.2022.3143566. URL https://ieeexplore.ieee.org/abstract/document/9684689?casa_token=3jWK_3wYpwYAAAAA:LQ8E6Ut5GswY1aLZdf6DF1sC4y4CtosaRQMOjRj19lQ0Ax9DG7E Conference Name: IEEE Transactions on Dependable and Secure Computing.

[19] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, April 2015. URL http://arxiv.org/abs/1412.1897. arXiv:1412.1897 [cs].

[20] Yaru Niu, Rohan Paleja, and Matthew Gombolay. Multi-Agent Graph-Attention Communication and Teaming. 2021.

[21] Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. SpringerBriefs in Intelligent Systems. Springer International Publishing, Cham, 2016. ISBN 978-3-319-28927-4 978-3-319-28929-8. doi: 10.1007/978-3-319-28929-8. URL http://link.springer.com/10.1007/978-3-319-28929-8.

[22] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks against Machine Learning, March 2017. URL http://arxiv.org/abs/1602.02697. arXiv:1602.02697 [cs].

[23] Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust Deep Reinforcement Learning with Adversarial Attacks, December 2017. URL http://arxiv.org/abs/1712.03632. arXiv:1712.03632 [cs].

[24] Neil C. Rabinowitz, Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine Theory of Mind, March 2018. URL http://arxiv.org/abs/1802.07740. arXiv:1802.07740 [cs].

[25] Melanie Sclar, Graham Neubig, and Yonatan Bisk. Symmetric Machine Theory of Mind. In *Proceedings of the 39th International Conference on Machine Learning*, pages 19450–19466. PMLR, June 2022. URL https://proceedings.mlr.press/v162/sclar22a.html. ISSN: 2640-3498.

[26] Tianmin Shu and Yuandong Tian. M$ˆ3$RL: Mind-aware Multi-agent Management Reinforcement Learning, March 2019. URL http://arxiv.org/abs/1810.00147. arXiv:1810.00147 [cs, stat].

[27] Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin A. Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua B. Tenenbaum, and Tomer D. Ullman. AGENT: A Benchmark for Core Psychological Reasoning, July 2021. URL http://arxiv.org/abs/2102.12321. arXiv:2102.12321 [cs].

[28] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. Learning when to Communicate at Scale in Multiagent Cooperative and Competitive Tasks, December 2018. URL http://arxiv.org/abs/1812.09755. arXiv:1812.09755 [cs, stat].

[29] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning Multiagent Communication with Backpropagation, October 2016. URL http://arxiv.org/abs/1605.07736. arXiv:1605.07736 [cs].

[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, February 2014. URL http://arxiv.org/abs/1312.6199. arXiv:1312.6199 [cs].

[31] James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Adversarial Attacks On Multi-Agent Communication, October 2021. URL http://arxiv.org/abs/2101.06560. arXiv:2101.06560 [cs].

[32] Rundong Wang, Xu He, Runsheng Yu, Wei Qiu, Bo An, and Zinovi Rabinovich. Learning Efficient Multi-agent Communication: An Information Bottleneck Approach, June 2020. URL http://arxiv.org/abs/1911.06992. arXiv:1911.06992 [cs].

[33] Yuanfei Wang, Fangwei Zhong, Jing Xu, and Yizhou Wang. ToM2C: Target-oriented Multi-agent Communication and Cooperation with Theory of Mind, April 2022. URL http://arxiv.org/abs/2111.09189. arXiv:2111.09189 [cs].

[34] Hao Xiang, Runsheng Xu, Xin Xia, Zhaoliang Zheng, Bolei Zhou, and Jiaqi Ma. V2XP-ASG: Generating Adversarial Scenes for Vehicle-to-Everything Perception. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3584–3591, May 2023. doi: 10.1109/ICRA48891.2023.10161384. URL https://ieeexplore.ieee.org/abstract/document/10161384.

[35] Wanqi Xue, Wei Qiu, Bo An, Zinovi Rabinovich, Svetlana Obraztsova, and Chai Kiat Yeo. Mis-spoke or mis-lead: Achieving Robustness in Multi-Agent Communicative Reinforcement Learning, January 2022. URL http://arxiv.org/abs/2108.03803. arXiv:2108.03803 [cs].

[36] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations, July 2021. URL http://arxiv.org/abs/2003.08938. arXiv:2003.08938 [cs, stat].

[37] Changxi Zhu, Mehdi Dastani, and Shihan Wang. A Survey of Multi-Agent Reinforcement Learning with Communication, March 2022. URL http://arxiv.org/abs/2203.08975. arXiv:2203.08975 [cs].