MEDIATE: Mutually Endorsed Distributed Incentive Acknowledgment Token Exchange

Philipp Altmann[1], Katharina Winter[2], Michael Kölle[1], Maximilian Zorn[1] and Claudia Linnhoff-Popien[1]

[1]*LMU Munich, Germany*

[2]*Munich University of Applied Sciences, Munich, Germany*

Keywords: Multi-Agent Systems, Reinforcement Learning, Peer Incentivization, Consensus, Emergent Cooperation.

Abstract: Recent advances in *multi-agent systems* (MAS) have shown that incorporating *peer incentivization* (PI) mechanisms vastly improves cooperation. Especially in social dilemmas, communication between the agents helps to overcome sub-optimal Nash equilibria. However, incentivization tokens need to be carefully selected. Furthermore, real-world applications might yield increased privacy requirements and limited exchange. Therefore, we extend the PI protocol for *mutual acknowledgment token exchange* (MATE) and provide additional analysis on the impact of the chosen tokens. Building upon those insights, we propose *mutually endorsed distributed incentive acknowledgment token exchange* (MEDIATE), an extended PI architecture employing automatic token derivation via decentralized consensus. Empirical results show the stable agreement on appropriate tokens yielding superior performance compared to static tokens and state-of-the-art approaches in different social dilemma environments with various reward distributions.

1 INTRODUCTION

Recent advances in using reinforcement learning (RL) in multi-agent systems (MAS) demonstrated their feasibility for real-world multi-agent reinforcement learning (MARL) applications. Those applications range from smart grids (Omitaomu and Niu, 2021) and factories (Kim et al., 2020) to intelligent transportation systems (Qureshi and Abdullah, 2013). To assess the agents' cooperation capabilities, social dilemmas producing tensions between the individual and collective reward maximization (social welfare) are often used (Dawes, 1980). Yet, the availability of communication and exchange is vital to fostering cooperation between self-interested individuals. However, besides the autonomous interaction within an environment, increased privacy requirements might require instances to conceal information regarding their current state (Tawalbeh et al., 2020). Peer incentivization (PI) is a recent branch of research offering a distinct solution for emergent cooperation between agents. At its core, PI enables agents to shape each other's behavior by exchanging reward tokens in addition to the environmental reward (Phan et al., 2022; Lupu and Precup, 2020). However, for proper integration and effective incentivization, those

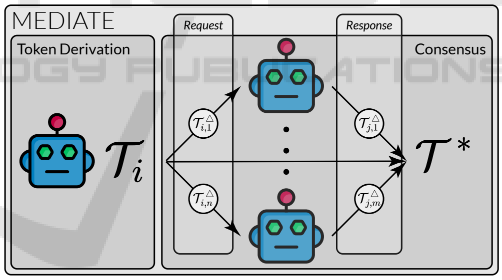


Figure 1: MEDIATE Architecture deriving a consensual PI token $\mathcal{T}^*$ through reciprocal decentralized communication.

exchanged tokens need to be carefully considered, regardless of whether their value is set dynamic or as a hyperparameter. For the robust and scalable applicability of PI mechanisms in decentralized learning scenarios, adaptive incentivization tokens and mechanisms to agree upon common token values are required. Yet, current approaches are missing said coordinated adaptability. To overcome these shortcomings, we provide the following contributions:

- We evaluate the effect of different centralized (common) and decentralized (varying) values for the incentivization token.
- We propose *mutually endorsed distributed incentive acknowledgment token exchange* (MEDIATE,

Altmann, P., Winter, K., Kölle, M., Zorn, M. and Linnhoff-Popien, C.
MEDIATE: Mutually Endorsed Distributed Incentive Acknowledgment Token Exchange.
DOI: 10.5220/0013091900003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 1*, pages 33-44
ISBN: 978-989-758-737-5; ISSN: 2184-433X

cf. Fig. 1), an automatic token derivation mechanism based on the agents' value estimate, and a consensus mechanism to mediate a global token maintaining local privacy.

- We provide ablation studies of the introduced token derivation and the consensus mechanism over a static token. Benchmark comparisons to state-of-the-art PI approaches show that MEDIATE can negotiate appropriate tokens that yield improved cooperation and social welfare in various social dilemmas with different reward landscapes.

2 PRELIMINARIES

Social Dilemmas. Game Theory analyzes behavior among rational agents in cooperative and competitive situations (Russell, 2010; Littman, 2001). Social dilemmas are Markov games that inhibit a specific reward structure, which creates tension between individual and collective reward maximization. *Sequential social dilemmas* (SSD) are temporally extended social dilemmas, in which the game repeats over several time steps (Leibo et al., 2017). The Nash equilibrium is a situation where no agent can increase its individual reward by changing its strategy if all other agents maintain their current strategy (Littman, 2001; Sandholm and Crites, 1996). MARL utilizes SSDs to analyze and experiment with the social behavior of different learning strategies (Leibo et al., 2017). To assess the emergence of cooperation, we employ the *Iterated Prisoner's Dilemma* (IPD), where mutual defection constitutes a Nash equilibrium (Axelrod, 1980; Sandholm and Crites, 1996). To evaluate the scalability of our approach, we use the *Coin Game* with two, four, and six agents (Lerer and Peysakhovich, 2017). Additionally, we use the *Rescaled Coin Game* with two agents to assess the robustness w.r.t. varying reward landscapes. The rate of *own coins* versus total coins collected reflects overall cooperation. For insights on long-term cooperation, we use *Harvest*, posing a risk of the *tragedy of the commons* to self-interested agents (Perolat et al., 2017; Phan et al., 2022). For further details about the environments used, please refer to the Appendix.

Peer Incentivization. In MAS, cooperation connotes the joining of individual problem-solving strategies of autonomous agents into a combined strategy (Crainic and Toulouse, 2007). The emergent cooperation of learning agents necessitates coordination (Noë, 2006), which poses a vital challenge to current communication protocols in decentralized MARL scenarios (Jaques et al., 2019; Kölle et al., 2023; Altmann et al., 2024b). PI is a recent branch of research, focussing on agents learning to actively shape the behavior of others by sending rewards or penalties (Phan et al., 2022; Yang et al., 2020). These peer rewards are processed like environment rewards, enabling the emergence of cooperation. However, new dynamics arise through the increased inter-dependency, which comes with new challenges. Carefully designing this reward mechanism is essential to achieving a good outcome (Lupu and Precup, 2020).

Consensus in Multi-Agent Systems. Distributed systems use consensus algorithms to deduct a global average of local information (Schenato and Gamba, 2007). For MAS, consensus describes the convergence of agents on a mutual value via communication (Li and Tan, 2019). A consensus algorithm specifies the execution steps to reach consensus (Han et al., 2013). Bee swarms, bird flocks, and other group-coordinated species show natural behavior (Amirkhani and Barshooi, 2022) that inspires further underlying concepts like leadership, voting, or decision-making (Conradt and Roper, 2005). Two main application areas for consensus algorithms are sensor networks (Yu et al., 2009) and blockchain technology (Monrat et al., 2019), which has played an integral role in cryptocurrencies and provides promising solutions for IoT applications. Consensus in sensor networks mainly deals with the fusion of distributed data, especially for time-critical data (Schenato and Gamba, 2007) and uncertainty in large-scale networks (Olfati-Saber and Shamma, 2005). Research in cryptocurrency and IoT focuses on synchronization (Cao et al., 2019), agreement (Salimitari and Chatterjee, 2018), and verification of actions (Lashkari and Musilek, 2021) between entities in distributed systems. The number of sophisticated consensus algorithms is growing through the rising importance of decentralized coordination mechanisms (Lashkari and Musilek, 2021) in an increasingly digitally connected world. Our approach utilizes the cryptographic technique of additive secret sharing, solving the average consensus problem for privacy-critical tasks (Li et al., 2019). MARL research on consensus algorithms has been increasing recently, intending to reach an optimal joint policy in a decentralized system that is robust to unreliable agents or adversarial attacks (Figura et al., 2021). To our knowledge, no research exists concerning consensus algorithms, PI and RL.

Problem Formulation. We formulate our problem of a MAS as a *stochastic game* $\mathcal{M} = \langle \mathcal{D}, \mathcal{S}, \mathcal{Z}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$, with the set of all agents $\mathcal{D} =$

$\{1, \ldots, N\}$, a set $\mathcal{S}$ of states s_t at time step t, a set $\mathcal{A} = \langle \mathcal{A}_1, \ldots, \mathcal{A}_N \rangle$ of joint actions $a_t = \langle a_{t,i} \rangle_{i \in \mathcal{D}}$, the transition probability $\mathcal{P}(s_{t+1} \mid s_t, a_t)$, and the joint reward $\mathcal{R}(s_t, a_t) = \langle r_{t,i} \rangle_{i \in \mathcal{D}} \in \mathbb{R}$. Furthermore, we assume each agent i to have a neighborhood $\mathcal{N}_{t,i} \subseteq \mathcal{D} \setminus \{i\}$, bounding its set of local observations $z_{t+1} = \langle z_{t+1,i} \rangle_{i \in \mathcal{D}} \in \mathcal{Z}^N$, and the agents' experience tuple $\langle \tau_{t,i}, a_{t,i}, r_{t,i}, z_{t+1,i} \rangle$, where $\tau_{t,i} \in (\mathcal{Z} \times \mathcal{A}_i)_t$ is the agent's history. Agent i selects the next action based on a stochastic policy $\pi_i(a_{t,i} | \tau_{t,i})$. Simultaneously learning agents cause non-stationary, i.e., varying transition probabilities over time. The goal of each self-interested agent i is to find a *best response* π_i^* that maximizes the expected individual discounted return:

$$G_{t,i} = \sum_{k=0}^{\infty} \gamma^k r_{i,t+k}, \tag{1}$$

with a discount factor $\gamma \in [0, 1)$. From the perspective of an agent, other agents are part of its environment, and policy updates by other agents affect the performance of an agent's own policy (Laurent et al., 2011). The performance of π_i is evaluated using a *value function* $V_i(s_t) = \mathbb{E}_\pi[G_{t,i} | s_t]$ for all $s_t \in \mathcal{S}$, with the *joint policy* $\pi = \langle \pi_j \rangle_{j \in \mathcal{D}}$ (Buşoniu et al., 2010). Both the policies π and the value functions V are approximated by independent neural networks parameterized by θ and ω, respectively. For simplicity, we omit those for the following and use the abbreviated forms $V_i = V_i^\omega(\tau_{t,i}) \approx V^{\pi_i}(s_t)$ and $\pi_i = \pi_i^\theta$ respectively. To measure *efficiency* U of the whole MAS, we furthermore consider the social welfare (Sandholm and Crites, 1996), measured by the sum of undiscounted returns over all agents within an episode until time step T:

$$U = \sum_{i \in \mathcal{D}} \sum_{t=0}^{T-1} r_{t,i} \tag{2}$$

Furthermore, we use the fraction of *own coins* to measure cooperation based on the coins collected by each agent:

$$\texttt{own\_coins} = \frac{\#\,\texttt{own coins collected}}{\#\,\texttt{total coins collected}} \tag{3}$$

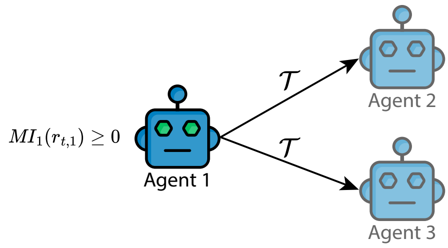
Mutual Acknowledgment Token Exchange (MATE). MATE is a reciprocal approach to PI based on a two-phase communication protocol, as shown in Fig. 2, to exchange *acknowledgment tokens* $\mathcal{T} \geq 0$ for individual reward shaping of $r_{t,i}$, depending on a *monotonic improvement measure* MI_i. $MI_i(\overline{r}_{t,i})$ is defined by the temporal difference residual of $\hat{V}_i$ w.r.t. some arbitrary reward $\overline{r}_{t,i}$ as proposed in (Phan et al., 2022, 2024):

$$MI_i(\overline{r}_{t,i}) = \overline{r}_{t,i} + \gamma \hat{V}_i(\tau_{t+1,i}) - \hat{V}_i(\tau_{t,i}) \tag{4}$$

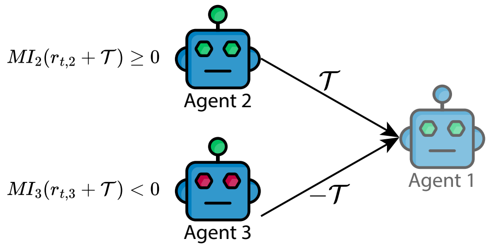
In the *request phase* (Fig. 2a), each agent i checks its current situation via MI_i. If $MI_i(r_{t,i}) \geq 0$, the agent sends a *token* $x_i = \mathcal{T}$ as an *acknowledgment request* to all other agents $j \in \mathcal{N}_{t,i}$ as a reward. In the *response phase* (Fig. 2b), all request-receiving agents $j \in \mathcal{N}_{t,i}$ check if the request token x_i would improve their situation along with their own respective reward $r_{t,j}$. If $MI_j(r_{t,j} + x_i) \geq 0$, then agent j accepts the request with a positive *response token* $y_j = \mathcal{T}$. However if $MI_j(r_{t,j} + x_i) < 0$, then agent j rejects the request with a negative response token $y_j = -\mathcal{T}$. After the request and response phase, the shaped MATE reward is computed for each agent i as follows:

$$\hat{r}_{t,i} = r_{t,i} + max\{\langle x_j \rangle_{j \in \mathcal{N}_{t,i}}\} + min\{\langle y_j \rangle_{j \in \mathcal{N}_{t,i}}\} \tag{5}$$

In the following, we will use the MATE protocol (Fig. 2) and reward (Eq. 5) without any change and explain our contributions on top of it.

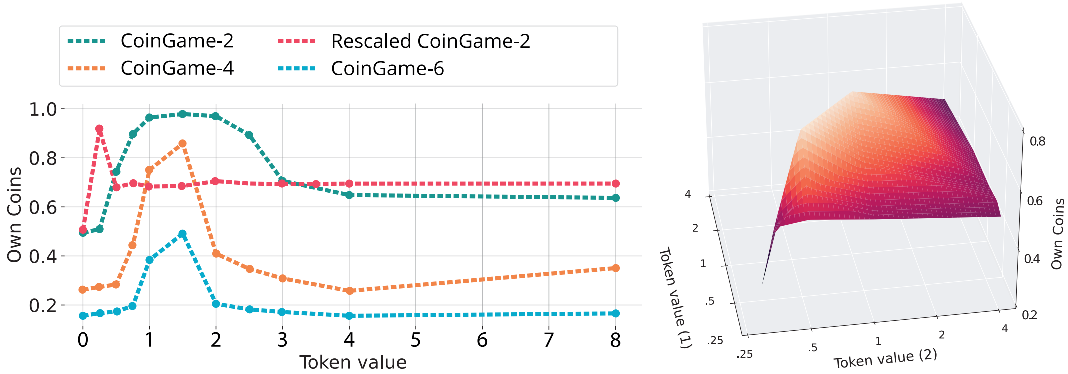


(a) Request phase



(b) Response phase

Figure 2: MATE protocol example. (a) If agent 1 estimates a monotonic improvement $MI_1(r_{t,1}) \geq 0$ of its situation, it "thanks" its neighbor agents 2 and 3 by sending an *acknowledgment request* $\mathcal{T}$ as reward. (b) Agent 2 and 3 check if the request $\mathcal{T}$ monotonically improves their own situation along with their own respective reward. If so, a positive reward $\mathcal{T}$ is sent back as a response. If not, a negative reward $-\mathcal{T}$ is sent back.



(a) Central token values for CoinGame-2, -4, and -6 (b) Decentralized Token Values for CG-2

Figure 3: Rate of *own coins* for different tokens when determined centralized (3a) and decentralized (3b).

3 IMPACT OF INCENTIVIZATION TOKENS

As MATE was previously only evaluated with token values of 1, we first aim to provide additional insights into the impact of the incentivization token, supplying an extensive hyperparameter analysis, both per-agent (*decentralized*, i.e., x_i and y_j may differ) and globally (*centralized*, i.e., $x_i = y_j = \mathcal{T}$). Fig. 3a displays the level of cooperation measured by the rate of *own coins* collected for different token values $\mathcal{T} \in [0, 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, 3, 4, 8]$ in the Coin Game with two, four, and six agents, as well as the two-agent Coin Game with scaled rewards. We averaged all results over five random seeds. The graphs display high average levels of cooperation for value 1 in all settings, except for the down-scaled Coin Game, where token 1 fails. This indicates that the token value is highly dependent on the reward landscape. Insufficient (inferior) token values fail to achieve the collective objective, causing self-interested behavior. Conversely, over-exploitative (intemperate) token values likewise fail to yield cooperative behavior. As the number of agents increases, a value of 1.5 appears optimal within the presented range, but the required precision for successful cooperation varies. Also, the range of token values that yield high cooperation narrows, retaining its relative position but exhibiting an increased sensitivity to the boundaries of that range. The discrepancy between the optimal token value of 1.5 and the value of 1 increases in the six-agent Coin Game.

The analysis implies that factors like the domain, the reward landscape, and the number of agents influence incentive rewards. The range of tokens with distinctively high cooperation is solely a function of the environment rewards but depends on the specific dynamics of the game, making it challenging to predict. A fixed token value lacks the adaptability required for diverse settings, making a priori prediction based on parameter settings a complex task. It becomes evident that reward structures are not the sole determinants for selecting appropriate token weights and may not even be reliably indicative across all scenarios.

To provide further insights into the dynamics introduced by the choice of incentivization token value, we modified the protocol to allow the agents to exchange disparate tokens. We refer to this mode as *decentralized*. Note that using automated token derivation in a decentralized setting without a mechanism for coordination or consensus might result in such varying token values. Fig. 3b maps the interpolated cooperation levels in the two-agent Coin Game with the tokens $\mathcal{T} \in [0.25, 0.5, 1, 2, 4]$, as values between 1 and 2 have previously shown to be sufficient central tokens, employed by both agents, measured by the rate of own coins. The results reveal that the token combinations $(1, 1)$ and $(2, 2)$ yield the highest cooperation rates. Both token values are positioned in the appropriate token range in the centralized comparison (cf. Fig. 3a), and the combinations contain equal values, which appears to be a significant criterion in this context. Although the combination $(1, 2)$ includes two appropriate values, the cooperation is decreased compared to the equal-valued exchange. With increasing discrepancy between the token values, cooperation further decreases, suggesting a correlation between the degree of value equality and cooperation. Agents with over-exploitative token values can greatly impact other agents, especially those with limited social influence due to smaller tokens, leading to a manipulative form of cooperation. Equal but inappropriate token values exhibit low performance and cooperation, which minimizes for $(0.25, 0.25)$.

Overall, this evaluation suggests that the exchange of decentralized token values must be appropriate and equal to provide fairness and induce equal cooperation. Nevertheless, the rate of own coins collected for all tested tokens excels the performance of naïve learning, reflected by token value 0. Conceptually, these prospects of MATE arise from enabling agents to share their success, provided the benefits are mutual. As shown before, however, exchanging tokens of value $\mathcal{T} = 1$ might not always be a sufficient choice for any given environment.

4 MEDIATE

To elevate PI token values from static hyperparameters to dynamically adaptable domain-specific quantities, we propose *mutually endorsed distributed incentive acknowledgment token exchange* (MEDIATE), combining two progressions (cf. Fig. 1): First, we provide an automated mechanism to derive dynamic agent-based incentivization tokens $\mathcal{T}_i$. To ensure global convergence of said tokens, we secondly provide a consensus mechanism that ensures the privacy of the agents' local information.

Generally, we intend to improve cooperation by introducing reciprocal participation (via a positive reward or incentive) if agents experience monotonic improvement, i.e., their experienced situation is better than approximated by their local value estimate, causing a positive temporal difference. By requiring mutual acknowledgment of this improvement, convergence towards a strategy maximizing efficiency or social welfare is attained. Thus, mutual PI acts similarly to a global value function regularizing policy updates. To further support this effect, we retrieve the dynamic token values based on the agents' local value V_i. This allows us to provide a lightweight extension, not relying on additional models to be learned (in contrast to previous automatic incentivization approaches). MEDIATE operates decentralized, individually calculating a token value for each agent based on their respective value functions. As these token values are directly used to shape the agents' reward (c.f. Eq. 5), incentivization is relative to the agents' value, pushing their strategies towards global cooperation (similar to the monotonic improvement) while maintaining *value privacy* (i.e., an agent does not know the value function of other agents). This assumption ensures both independence and decentralization by enabling an agent to operate solely based on its domain-specific metrics and variables. Alg. 1 depicts the proposed mechanism for deriving and updating individual tokens.

Algorithm 1: Agent-wise Token Derivation with MEDIATE.

Setup for Agent $i \in \mathcal{D}$: $\mathcal{T}_i \leftarrow 0.1; r_i^{\min} \leftarrow \infty; \widetilde{V}_i \leftarrow 0$

for Epoch ϵ in Epochs; Agent $i \in \mathcal{D}$ **do**
  $\overline{V}_i \leftarrow \{\}$ ▷ *Initialize mean values for epoch*
  for Rollout $\langle \tau_{0,i}, a_{0,i}, r_{0,i}, \ldots, \tau_{T,i}, a_{T,i}, r_{T,i} \rangle$ in ϵ **do**
    $r_i^{\min} \leftarrow \min(r_i^{\min}, \langle r_{0 \ldots T,i} \rangle)$
    $\overline{V}_i \leftarrow \overline{V}_i \cup \bar{V}_i(\tau)$ ▷ *Calculate mean value* (6)
  end for
  $\mathcal{T}_i \leftarrow \max(\mathcal{T}_{(i)}^{(*)} + \nabla_{\mathcal{T}_i}, 0)$ ▷ *Update local token* (7)
  $\widetilde{V}_i \leftarrow \text{median}(\overline{V}_i)$
end for

All agents initially set their token to a small but non-zero value of 0.1 to differentiate it from a zero-valued token that would equate to naïve learning. This initialization allows for the immediate incorporation of the PI mechanism. To ensure an appropriate acceptance-rejection-ratio and thus an appropriate impact on the behavior of other agents, the token value must be proportional to the value function. Thus, we suggest incrementing tokens by the relative difference between the mean state value estimates across consecutive epochs. By doing so, MEDIATE tailors tokens to the unique dynamics of each domain, thereby fostering equal cooperation across diverse settings. As a measure of the profit, we derive the mean accumulated value $\bar{V}$ of an episode τ of length T similar to the undiscounted return (cf. Eq. (1)):

$$\bar{V}_i(\tau) = \frac{\sum_{t=0}^{T} V_i(\tau_{t,i})}{T} \tag{6}$$

V_i refers to the current value approximation of agent i. Furthermore, we use the median of the mean values $\bar{V}$ over an epoch of episodes to improve stability. The local tokens $\mathcal{T}_i$ are adjusted every epoch based on the difference (Δ) between the current median of the mean values (median($\overline{V}_i$)) and the previous median of the mean value $\widetilde{V}_i$:

$$\nabla_{\mathcal{T}_i} = \alpha \cdot \frac{\Delta(\widetilde{V}_i, \text{median}(\overline{V}_i))}{\widetilde{V}_i} \cdot |r_i^{\min}|, \tag{7}$$

with $\alpha = 0.1$ as a constant comparable to a learning rate and the absolute value of the lowest encountered environmental reward $r_i^{\min}$ (cf. Alg. 1) as a scaling factor. Furthermore, we use the previous median of the mean value $\widetilde{V}_i$ for normalization. Consequently, sufficiently large negative state value estimates can cause positive tokens, which rise when the value further decreases. For negative values, the token thus remains proportionate to the absolute magnitude of the value function. Furthermore, the resulting token value is clamped to positive values using the $\max$ operation (cf. Alg. 1), sending a zero token otherwise. Resembling the use of a ReLU activation function (Agarap, 2018), this forces the agent to send no incentive when unable to send a positive. By this, agents adhere to the principle of *Niceness*, which is a core principle for the reciprocal strategy of MATE, implying no intent of defection in the request (Phan et al., 2022).

However, besides using appropriate tokens, findings from the analysis of decentralized tokens also demonstrated the need for equal token values in the mutual exchange. Therefore, we extend MEDIATE with a consensus mechanism to reach an agreement on a mutual token, increasing equality and reducing the impact of outliers while preserving the privacy of the agents' confidential information using additive secret sharing. All agents set up the consensus exchange by dividing their token values into shares for all agents in their neighborhood $\mathcal{N}$, reserving one share for privacy reasons. The token is only reconstructable when accounting for all shares, which provides security against privacy defectors. In the request phase, all agents i send the corresponding shares $[\mathcal{T}_{i,1}^{\triangle}, \ldots, \mathcal{T}_{i,n}^{\triangle}]$ to all n neighbors. Each receiving agent j accumulates its received shares $[\mathcal{T}_{j,1}^{\triangle}, \ldots, \mathcal{T}_{j,m+1}^{\triangle}]$ from its m neighbors, including its reserved share. In the response phase, each agent j sends the accumulated shares to all its neighbors. Each receiving agent i obtains the accumulated shares from all neighbors, which it averages over the number of shares, i.e., the number of agents N, to obtain the reconstructed consensus token $\mathcal{T}^*$:

$$\mathcal{T}^* = \frac{\sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} \mathcal{T}_{i,j}^{\triangle}}{N} \tag{8}$$

In domains like Harvest, with only partially connected agents and changing topologies, the consensus protocol includes a multi-iteration response phase. Each summed share is tagged with an ID, sent to all neighbors, and forwarded over multiple time steps to ensure network-wide information dissemination. To integrate the reconstructed token into the token derivation mechanism, we propose two different update mechanisms: *Isolated* updates the local token $\mathcal{T}_i$ based on the previous local token, which is shared independently via the consensus protocol: $\max(\mathcal{T}_i + \nabla_{\mathcal{T}_i}, 0)$. In contrast, *synchronized* replaces the local token with the reconstructed token $\mathcal{T}^*$ after the consensus phase: $\max(\mathcal{T}^* + \nabla_{\mathcal{T}_i}, 0)$. Consequently, only the token update (cf. Alg. 1) is affected, either synchronized with the consensus token $\mathcal{T}^*$ or drifting independently. We will refer to the resulting variants as *MEDIATE-I* and *MEDIATE-S*.

5 RELATED WORK

Various concepts help achieve emergent cooperation in MAS. *Learning with opponent-learning awareness* (LOLA) (Foerster et al., 2018) and *stable opponent shaping* (SOS) (Letcher et al., 2019) consider the learning process of other agents and shape the policy updates of opponents. Nature and human social behavior also inspired many concepts. Wang et al. (2019) developed an evolutionary approach to create agents with social behavior by natural selection. Other work focuses on prosocial agents and intrinsic motivation thriving for the manifestation of social norms (Jaques et al., 2019). Eccles et al. (2019) divided agents into innovators, learning a policy, and imitators, which reciprocate innovators. Baumann et al. (2020) insert an external planning agent into the environment, which can observe all agents and distribute rewards. Overall, we divide approaches fostering emergent cooperation into constructed artificial social assemblies, added intrinsic motivation, and external optimization techniques. Our approach combines those concepts, using socially inspired mutual acknowledgment to shape the environmental rewards.

A large corpus in PI research focuses on similar approaches to learning incentives integrated into the model. *Gifting* integrates the reward-gifting capability into agents' policies as an additional action. Different reward mechanisms can build upon this concept. In *zero-sum gifting*, agents receive a penalty for each sent reward to balance the total sum of rewards. Gifting can also be only allowed up to a *fixed budget* per episode as an alternative to penalization. With a *replenishable budget*, the reception of environment rewards can recharge this budget (Lupu and Precup, 2020). *Learning to incentivize other learning agents* (LIO) is another approach that uses an incentive function to learn appropriate peer rewards. Selecting a reward is not part of the action space but is learned separately by a second model (Yang et al., 2020). Like LIO, MEDIATE derives incentives from the agents' expected environmental return. However, in contrast to MEDIATE, LIO requires an additional model to be learned to predict this value, which causes additional overhead. *Learning to share* (LToS) also implements two policies, one for local objectives set by a high-level policy (Yi et al., 2021). *Peer-evaluation-based dual-DQN* (PED-DQN) lets agents evaluate their received peer signals w.r.t. their environment rewards with an additional DQN network (Hostallero et al., 2020). *Learning to influence through evaluative feedback* (LIEF) learns to reconstruct the reward function of peers via feedback. The authors call for an investigation between a manual, systematic, and learned construction of rewards (Merhej and Chetouani, 2021). Fayad and Ibrahim (2021) use counterfactual simulations to derive influential actions. The above concepts modify the agent models or the action space to derive the intrinsic rewards. Rather than altering the agents themselves, we utilize an additional protocol layer, which serves as a tool for agents and yields increased flexibility.

Building upon *mutual acknowledgment token exchange* (MATE) (Phan et al., 2022, 2024), we control the exchange of incentives via a two-phase communication protocol (c.f. Fig. 2). In the request phase of each time step, all agents evaluate their *monotonic improvement* (MI), c.f. Eq. (4), and potentially send acknowledgment tokens to all neighbors. In the response phase, agents evaluate their MI w.r.t. the sum of environment rewards and the received token, and respond with a positive or negative token. This two-way handshake allows agents to give feedback to other agents when incentives are received, which fosters cooperation and has been shown to outperform naïve learning and other PI approaches, like LIO and Gifting, in various benchmarks regarding efficiency and equality metrics (Phan et al., 2022, 2024). MATE uses a communication layer and thus provides a lightweight solution with minimal interference with the agent model. Due to this flexible and privacy-conserving design, we evaluate our approach as an extension of MATE. However, note that other protocol PI solutions can also utilize MEDIATE.

Overall, we aim to eliminate the need to set the exchange token beforehand, which is a central limitation of MATE. Given their direct combination with the external reward, we argue that incentivization tokens are sensitive parameters to be carefully considered. Kuhnle et al. (2023) analyze the Harsanyi-Shapley value to determine the weight of a side payment based on the strategic strength of a player in two-player scenarios. Value decomposition networks (Sunehag et al., 2018), VAST (Phan et al., 2021), and QMIX (Rashid et al., 2020) decompose the joint action-value function into agent-based value functions to achieve cooperation and maximize social welfare. These approaches are based on a centralized value function, whereas our work focuses on independent learners in a fully decentralized setting. MEDIATE also uses the value function to automatically derive token values to be mixed with the environmental reward, posing a lightweight and efficient solution.

6 EXPERIMENTAL RESULTS

To assess the effect of the introduced token derivation mechanism and the proposed consensus architecture, we ran evaluations comparing *isolated* and *synchronized* MEDIATE in the *IPD*, *CoinGame-2*, and *CoinGame-4*. As an additional ablation, we use a reduced version with only the automated decentralized token derivation (cf. Alg. 1) without any consensus mechanism, which we refer to as *AutoMATE*. Additionally, we compare the above to naïve learning and MATE with a fixed token of 1. We measure cooperation in all Coin Game environments by the ratio between *own coins collected* (`occ`) and *total coins collected* (`tcc`): *own coins* $= \frac{|\texttt{occ}|}{|\texttt{tcc}|}$. We compare the performance in the *IPD* and *Harvest* by the approaches' *efficiency* (cf. Eq. (2)) as a metric for social welfare. Additionally, we compare all MEDIATE ablations w.r.t. the convergence of their token value. To test the scalability of MEDIATE and its robustness to varying reward distributions, we provide further evaluations in the *Rescaled Coin Game*-2, *CoinGame-6*, and *Harvest*, including benchmark comparisons to zero-sum- and budget-gifting and LIO.

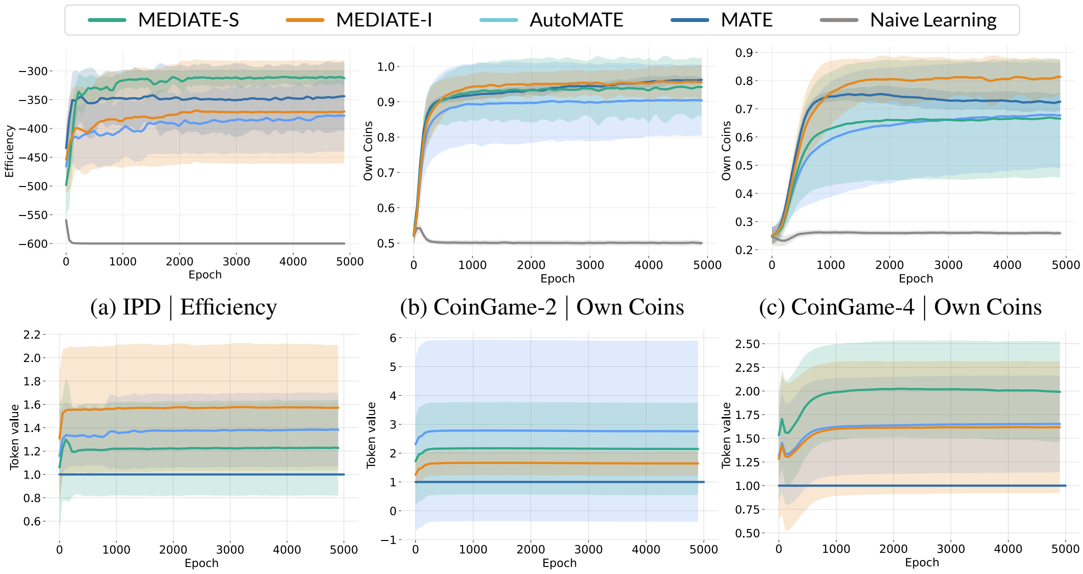
Training is conducted for 5000 epochs, comprising ten episodes each. We averaged all of the following results over eight random seeds. If not stated otherwise, all implementations use their default hyperparameters from the corresponding source. Please refer to the appendix for further environment- and implementation details [1].

6.1 Evaluation of MEDIATE

Fig. 4 shows the evaluation results. The graphs indicate that either synchronized or isolated MEDIATE updates consistently achieve efficiency and cooperation levels comparable or superior to MATE in all experimental settings, which legitimates their further investigation. As expected, naïve learning fails to reach emergent cooperation, again showcasing the compared environments' intricacy.

In general, MEDIATE enhances the performance of AutoMATE across all settings, except for the two-agent Coin Game scenario, where isolated updates neither improve nor deteriorate cooperation. The results imply that the combined automatic and decentralized mechanism - introduced by MEDIATE - provides sufficient tokens to replace the original MATE token value of 1. Furthermore, Figs. 4d-4f show that all automatically derived tokens converge within the initial 1000 epochs, indicating the purposeful nature

[1] All required implementations are available at https://github.com/philippaltmann/MEDIATE.



(a) IPD | Efficiency (b) CoinGame-2 | Own Coins (c) CoinGame-4 | Own Coins

(d) IPD | Token Values (e) CoinGame-2 | Token Values (f) CoinGame-4 | Token Values

Figure 4: MEDIATE Evaluation: Comparing the mean *Efficiency* (Fig. 4a) and rate of *Own Coins* (Fig. 4b, 4c) of *Naïve Learning* (grey), MATE (blue), AutoMATE (light blue), MEDIATE-I (orange), and MEDIATE-S (green), and the Mean *Token Value* (Fig. 4d, 4e, 4f) in the *IPD* (Fig. 4a, 4d), 2-agent *CoinGame* (Fig. 4b, 4e), and 4-agent *CoinGame* (Fig. 4c, 4f). The shaded areas mark the 95% confidence intervals. Overall, MEDIATE outperforms the compared approaches. Isolated consensus shows improved adaptability to increasing numbers of agents, while Synchronized consensus shows improved robustness in the negative-valued IPD.

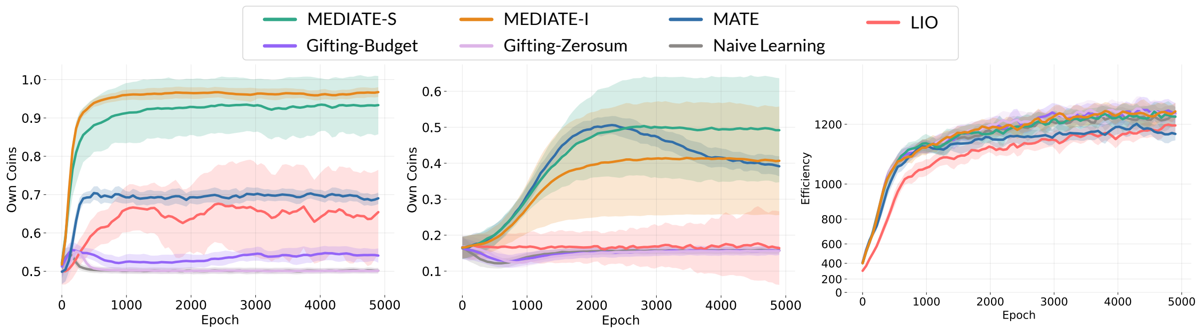
of the proposed architecture. In comparison, the corresponding tokens of AutoMATE and MEDIATE all converge to higher token values than MATE, which, according to our preliminary studies, are more optimal tokens. Wider confidence intervals in token convergence are generally associated with reduced efficiency and cooperation, but in the CoinGame-4, AutoMATE tokens converge to equivalent values as those with isolated updates. However, although its confidence interval is narrower, AutoMATE's performance is inferior due to the missing token coordination between the agents. Comparing the two MEDIATE variants, isolated updates perform better in both CoinGame settings.

In the negative-valued IPD domain, synchronized updates show advantages. Overall, in combination with the token plots, the results show that the update variant converging to a smaller value, i.e., the respectively less optimistic variant, provides superior tokens and thus yields improved efficiency and cooperation. Given the absence of a definitive superior option between the two MEDIATE variants, we include both in the benchmark comparisons.

6.2 Benchmark Comparisons

Fig. 5 shows the benchmark results. Table 1 summarizes the final performance metrics. The two-agent Coin Game features down-scaled rewards (RCG-2), requiring agents to learn cooperation under minimal positive and negative environment rewards. In contrast to the compared approaches, both MEDIATE variants achieve significantly higher rewards and master the task. Yet, isolated updates exhibit a slight performance advantage over synchronized updates. MATE demonstrates moderate cooperation, slightly improving upon LIO. In contrast, the gifting methods and naïve learning only show marginal cooperation, although Gifting-Budget performs comparably better. These results again highlight the superior adaptability of MEDIATE to unconventional, potentially challenging reward scenarios that yield improved applicability to varying tasks.

In the six-agent Coin Game (CG-6), naïve learning performs worst alongside Gifting-Zerosum and Gifting-Budget. While LIO shows a marginal improvement, it still lacks significantly behind MATE and MEDIATE regarding strategic cooperation. MEDIATE-I performs similarly to MATE, which potentially can be attributed to the limited capability of isolated updates to manage negative returns. MATE initially demonstrates an optimal learn-



(a) RCG-2 — Own Coins (b) CG-6 — Own Coins (c) Harvest (6 Agents) — Efficiency

Figure 5: Benchmark Comparison: Mean rate of *Own Coins* (Fig. 5a, 5b) and *Efficiency* (Fig. 5c) of MEDIATE-S (green), MEDIATE-I (orange), MATE (blue), LIO (red), *Budget-Gifting* (purple), *Zerosum-Gifting* (pink) and *Naïve Learning* (grey) in the *Rescaled CoinGame-2* (RCG-2) (Fig. 5a), *CoinGame-6* (CG-6) (Fig. 5b), and *Harvest* (Fig. 5c). The shaded areas mark the 95% confidence intervals. Across all scenarios, MEDIATE shows superior performance. Generally, using Isolated consensus shows increased adaptability to the intricate reward landscapes specifically considered here.

ing curve but deteriorates in performance afterward. In terms of cooperation, MEDIATE with synchronized updates emerges as performing best.

Harvest demonstrates the ability of MEDIATE to benefit in partially connected topologies. Here, MEDIATE ranks among the top-performing approaches and enhances the performance of MATE by providing an appropriate incentivization token. It thus demonstrates its efficacy in functioning even within unreliable environments while preserving privacy over the agents' local value information.

Table 1: Final average of the rate of *Own Coins* in the *Rescaled CoinGame-2* (RCG-2) and *CoinGame-6* (CG-6), and the *Efficiency* in *Harvest* for *synchronized* and *isolated* MEDIATE (MEDIATE-S, MEDIATE-I), AutoMATE, MATE, LIO, *Budget-* and *Zerosum* Gifting (Budget-G, Zerosum-G), and Naïve Learning.

	RCG-2	CG-6	Harvest
MEDIATE-S	0.93 ± 0.08	$\mathbf{0.50 \pm 0.16}$	1212 ± 20
MEDIATE-I	$\mathbf{0.97 \pm 0.02}$	0.41 ± 0.16	$\mathbf{1232 \pm 17}$
AutoMATE	0.86 ± 0.08	0.18 ± 0.09	1204 ± 35
MATE	0.69 ± 0.01	0.39 ± 0.03	*1177 ± 20*
LIO	0.69 ± 0.10	0.17 ± 0.11	1192 ± 20
Budget-G	0.54 ± 0.03	0.16 ± 0.02	1232 ± 23
Zerosum-G	*0.50 ± 0.01*	*0.16 ± 0.01*	1230 ± 20
Naïve L.	*0.50 ± 0.01*	*0.16 ± 0.01*	1220 ± 25

Overall, the evaluations demonstrated that emergent cooperation between agents fosters optimal social welfare. Appropriate reward weights can boost equal cooperation in social dilemmas, but such weights' appropriateness depends on the domain, the number of agents, the reward structure, or other factors. Involving a higher number of agents within a domain increases the required precision. Our experiments show that a token value of 1 - as proposed for MATE - is not universally appropriate in all domains or settings. In the down-scaled two-agent Coin Game, token value 1 is inappropriate, and in the six-agent Coin Game, it does not achieve optimal cooperation. Yet across all domains, MEDIATE exhibits strong adaptability while consistently delivering superior performance, even in challenging cooperative tasks such as the six-agent Coin Game, scenarios with complex reward landscapes, or unreliable environments with partially connected neighborhoods, like Harvest.

7 CONCLUSION

In this work, we proposed *mutually endorsed distributed incentive acknowledgment token exchange* (MEDIATE). MEDIATE introduces automated PI tokens in decentralized MAS with a consensus architecture and two agent-individual update mechanisms.

Token decentralization allows agents to use different tokens in the exchange. Experiments on the impact of different tokens in social dilemmas suggest that equal and appropriate token values foster improved social welfare. MEDIATE integrates the gradient of the agents' local value function approximation to derive appropriate tokens matching the external rewards. To achieve consensus on equal tokens, we propose extending the MATE protocol based on additive secret sharing, enabling the identification of the token average through the token exchange while adhering to privacy requirements. The consensus protocol is independent of the underlying algorithm for token derivation. We furthermore evaluate two token-update variations: A synchronized mechanism based on the reconstructed global token and an isolated mechanism using the previous local token.

Benchmark evaluations showed that MEDIATE achieves high social welfare in all tested domains. In all evaluated settings, MEDIATE improves the performance of MATE and even outperforms or matches the best-performing baselines. It represents a robust and adaptive solution capable of finding appropriate tokens. Computationally, MEDIATE is comparable to MATE while overcoming its central limitation of static token values. The only addition of deriving consented tokens at each update is a sum of constant values with linear complexity. Furthermore, the token extends on the value approximation. Thus, compared to LIO, no additional model needs to be learned.

Yet, even though not apparent in the evaluated social dilemma environments, this dependence on a robust value estimate also depicts a central limitation of MEDIATE. Therefore, integrating surrogate reward metrics like (Altmann et al., 2024a) might improve the overall robustness. Furthermore, the evaluated update mechanisms showed potentially unstable and prone to outliers. Thus, future work should focus on producing more accurate tokens, especially for an increased number of agents, making the overall algorithm more reliable in precision-requiring domains like the Rescaled CoinGame. Also, while MEDIATE has been shown to be robust to scaled reward landscapes, increasing numbers of agents, and long-term cooperation scenarios like Harvest, it should be tested for unreliable connections or defective scenarios.

Overall, MEDIATE provides a lightweight and robust framework to assess communication consensus mechanisms with automated peer incentives for emergent cooperation in various scenarios of social dilemmas.

ACKNOWLEDGEMENTS

This work is part of the Munich Quantum Valley, which is supported by the Bavarian state government with funds from the Hightech Agenda Bayern Plus.

REFERENCES

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Altmann, P., Ritz, F., Zorn, M., Kölle, M., Phan, T., Gabor, T., and Linnhoff-Popien, C. (2024a). Discriminative reward co-training. *Neural Computing and Applications*, pages 1–17.

Altmann, P., Schönberger, J., Illium, S., Zorn, M., Ritz, F., Haider, T., Burton, S., and Gabor, T. (2024b). Emergence in multi-agent systems: A safety perspective. In *Leveraging Applications of Formal Methods, Verification and Validation. Rigorous Engineering of Collective Adaptive Systems*, ISoLA '24, pages 104–120. Springer Nature.

Amirkhani, A. and Barshooi, A. H. (2022). Consensus in multi-agent systems: a review. *Artificial Intelligence Review*, 55(5):3897–3935.

Axelrod, R. (1980). Effective choice in the prisoner's dilemma. *Journal of conflict resolution*, 24(1):3–25.

Baumann, T., Graepel, T., and Shawe-Taylor, J. (2020). Adaptive mechanism design: Learning to promote cooperation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Buşoniu, L., Babuška, R., and De Schutter, B. (2010). Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1*, pages 183–221.

Cao, B., Li, Y., Zhang, L., Zhang, L., Mumtaz, S., Zhou, Z., and Peng, M. (2019). When internet of things meets blockchain: Challenges in distributed consensus. *IEEE Network*, 33(6):133–139.

Conradt, L. and Roper, T. J. (2005). Consensus decision making in animals. *Trends in ecology & evolution*, 20(8):449–456.

Crainic, T. G. and Toulouse, M. (2007). Explicit and emergent cooperation schemes for search algorithms. In *International Conference on Learning and Intelligent Optimization*, pages 95–109. Springer.

Dawes, R. M. (1980). Social dilemmas. *Annual review of psychology*, 31(1):169–193.

Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., and Leibo, J. Z. (2019). Learning reciprocity in complex sequential social dilemmas. *arXiv preprint arXiv:1903.08082*.

Fayad, A. and Ibrahim, M. (2021). Influence-based reinforcement learning for intrinsically-motivated agents. *arXiv preprint arXiv:2108.12581*.

Figura, M., Kosaraju, K. C., and Gupta, V. (2021). Adversarial attacks in consensus-based multi-agent reinforcement learning. In *2021 American Control Conference (ACC)*, pages 3050–3055. IEEE.

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2018). Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, page 122–130.

Han, Y., Lu, W., and Chen, T. (2013). Cluster consensus in discrete-time networks of multiagents with intercluster nonidentical inputs. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4):566–578.

Hostallero, D. E., Kim, D., Moon, S., Son, K., Kang, W. J., and Yi, Y. (2020). Inducing cooperation through reward reshaping based on peer evaluations in deep multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 520–528.

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J. Z., and De Freitas, N. (2019). Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Interna-*

tional conference on machine learning, pages 3040–3049. PMLR.

Kim, Y. G., Lee, S., Son, J., Bae, H., and Do Chung, B. (2020). Multi-agent system and reinforcement learning approach for distributed intelligence in a flexible smart manufacturing system. *Journal of Manufacturing Systems*, 57:440–450.

Kölle, M., Matheis, T., Altmann, P., and Schmid, K. (2023). Learning to participate through trading of reward shares. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, ICAART '23, pages 355–362. SciTePress.

Kuhnle, A., Richley, J., and Perez-Lavin, D. (2023). Learning strategic value and cooperation in multi-player stochastic games through side payments. *arXiv preprint arXiv:2303.05307*.

Lashkari, B. and Musilek, P. (2021). A comprehensive review of blockchain consensus mechanisms. *IEEE Access*, 9:43620–43652.

Laurent, G. J., Matignon, L., Fort-Piat, L., et al. (2011). The world of independent learners is not markovian. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 15(1):55–64.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473.

Lerer, A. and Peysakhovich, A. (2017). Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*.

Letcher, A., Foerster, J., Balduzzi, D., Rocktäschel, T., and Whiteson, S. (2019). Stable opponent shaping in differentiable games. In *International Conference on Learning Representations*.

Li, Q., Cascudo, I., and Christensen, M. G. (2019). Privacy-preserving distributed average consensus based on additive secret sharing. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE.

Li, Y. and Tan, C. (2019). A survey of the consensus for multi-agent systems. *Systems Science & Control Engineering*, 7:468 – 482.

Littman, M. L. (2001). Value-function reinforcement learning in markov games. *Cognitive systems research*, 2(1):55–66.

Lupu, A. and Precup, D. (2020). Gifting in multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on autonomous agents and multiagent systems*, pages 789–797.

Merhej, R. and Chetouani, M. (2021). Lief: Learning to influence through evaluative feedback. In *Adaptive and Learning Agents Workshop (AAMAS 2021)*.

Monrat, A. A., Schelén, O., and Andersson, K. (2019). A survey of blockchain from the perspectives of applications, challenges, and opportunities. *IEEE Access*, 7:117134–117151.

Noë, R. (2006). Cooperation experiments: coordination through communication versus acting apart together. *Animal behaviour*, 71(1):1–18.

Olfati-Saber, R. and Shamma, J. S. (2005). Consensus filters for sensor networks and distributed sensor fusion. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 6698–6703. IEEE.

Omitaomu, O. A. and Niu, H. (2021). Artificial intelligence techniques in smart grid: A survey. *Smart Cities*, 4(2):548–568.

Perolat, J., Leibo, J. Z., Zambaldi, V., Beattie, C., Tuyls, K., and Graepel, T. (2017). A multi-agent reinforcement learning model of common-pool resource appropriation.

Phan, T., Ritz, F., Belzner, L., Altmann, P., Gabor, T., and Linnhoff-Popien, C. (2021). Vast: Value function factorization with variable agent sub-teams. In *Advances in Neural Information Processing Systems*, NeurIPS '21, pages 24018–24032. Curran Associates, Inc.

Phan, T., Sommer, F., Altmann, P., Ritz, F., Belzner, L., and Linnhoff-Popien, C. (2022). Emergent cooperation from mutual acknowledgment exchange. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1047–1055.

Phan, T., Sommer, F., Ritz, F., Altmann, P., Nüßlein, J., Kölle, M., Belzner, L., and Linnhoff-Popien, C. (2024). Emergent cooperation from mutual acknowledgment exchange in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 38(34).

Qureshi, K. N. and Abdullah, A. H. (2013). A survey on intelligent transportation systems. *Middle-East Journal of Scientific Research*, 15(5):629–642.

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., and Whiteson, S. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research*, 21(1):7234–7284.

Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.

Salimitari, M. and Chatterjee, M. (2018). A survey on consensus protocols in blockchain for iot networks. *arXiv preprint arXiv:1809.05613*.

Sandholm, T. W. and Crites, R. H. (1996). Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37(1-2):147–166.

Schenato, L. and Gamba, G. (2007). A distributed consensus protocol for clock synchronization in wireless sensor network. In *2007 46th ieee conference on decision and control*, pages 2289–2294. IEEE.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., and Graepel, T. (2018). Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, page 2085–2087.

Tawalbeh, L., Muheidat, F., Tawalbeh, M., and Quwaider, M. (2020). Iot privacy and security: Challenges and solutions. *Applied Sciences*, 10(12):4102.

Wang, J. X., Hughes, E., Fernando, C., Czarnecki, W. M., Duéñez Guzmán, E. A., and Leibo, J. Z. (2019).

Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, page 683–692.

Yang, J., Li, A., Farajtabar, M., Sunehag, P., Hughes, E., and Zha, H. (2020). Learning to incentivize other learning agents. *Advances in Neural Information Processing Systems*, 33:15208–15219.

Yi, Y., Li, G., Wang, Y., and Lu, Z. (2021). Learning to share in multi-agent reinforcement learning. *arXiv preprint arXiv:2112.08702*.

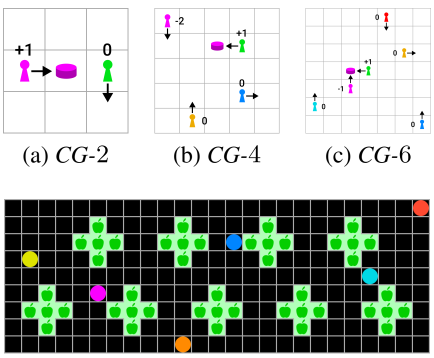
Yu, W., Chen, G., Wang, Z., and Yang, W. (2009). Distributed consensus filtering in sensor networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(6):1568–1577.

APPENDIX

Iterated Prisoner's Dilemma. The *Iterated Prisoner's Dilemma* (IPD) is the repeated game of the Prisoner's Dilemma, depicted in Table 2. At each time step, the two players must choose between cooperation and defection to maximize their payoff (Axelrod, 1980; Hostallero et al., 2020). Mutual defection constitutes a Nash equilibrium. If both agents defect, no agent is incentivized to change its strategy to cooperation in the next step if the other agent remains a defector. If both agents switched their strategy to cooperate, both would receive a lower penalty.

Table 2: Prisoner's Dilemma reward allocation. Each cell contains the respective payoffs for each of the two players based on their choice of cooperation or defection.

	Cooperate	**Defect**
Cooperate	(-1,-1)	(-3,0)
Defect	(0,-3)	(-2,-2)



(a) *CG*-2 (b) *CG*-4 (c) *CG*-6

(d) Harvest (6 agents)

Figure 6: Evaluation Environments

Coin Game. Coins or Coin Game is an SSD conceptualized by Lerer and Peysakhovich (2017). The *CoinGame*-N comprises $N \in \{2, 4, 6\}$ agents on a $3x3$, $5x5$, and $7x7$ grid respectively (cf. Figs. 6a-6b). A distinct color identifies each agent. Initially, all N agents and one random-colored coin spawn at random positions. The color of the coin matches one of the agents. An agent can distinguish whether the coin matches its own color or not. The action space of each agent comprises four directions of movement $\mathcal{A} \in \{left, right, up, down\}$. A coin is collected when an agent moves to its position. The environment discards actions violating its bounds. If an agent collects any coin, it receives a reward of $+1$. If the color matches a different agent, that agent is penalized with -2. If multiple agents collect a coin simultaneously, the matching agent receives a penalty of -1. Once a coin is collected, a new coin spawns. To evaluate varying reward scales, we added the *Rescaled Coin Game*-2 variation with downsized rewards (i.e., scaled by 0.1), such that the positive reward becomes $+0.1$ and the penalty weighs -0.2. The ratio between reward and penalty remains unchanged. Self-interested agents will collect all coins regardless of color since this strategy imposes only positive rewards on themselves. The Nash equilibrium is reached if all agents follow this strategy since refraining from collecting other agents' coins only reduces an agent's own rewards without mitigating the penalties incurred from the actions of other agents. However, if all agents collect their own coins, each agent profits from the reduced penalties, and social welfare can be maximized. To measure the level of strategic cooperation in this domain, we evaluate the rate of *own coins* w.r.t. to the total of collected coins.

Harvest. The Commons game is conceptualized by Perolat et al. (2017) and adapted by Phan et al. (2022), where it is named Harvest. In Harvest, agents move on a 25x9-sized grid to collect apples. The Harvest grid, including the fixed positioning of the apples, is displayed in Fig. 6d. Apples have a regrowth rate, which depends on the number of existing apples in the local area. More apples in the area cause a higher regrowth rate of collected apples. If no apples remain in the area, no apples regrow. Self-interested agents maximize their own apple harvest, but in a MAS, agents have to refrain from simultaneous apple collection to avoid the ultimate depletion of resources (the *tragedy of the commons*). This requirement is the Nash equilibrium of Harvest, as a single agent can not improve its rewards by refraining from apple collection when other agents will continue to diminish the resources. Only if all agents cooperate they can maximize their long-term rewards. Agents can tag other agents to remove them from the game for 25 time steps. (Perolat et al., 2017). In addition to a positive reward of $+1$ for an apple harvest, each time step poses a time penalty of -0.1. Furthermore, agents only have access to a partial observation surrounding their position. Agents can only communicate with agents in their neighborhood in an area of $7x7$ tiles. In addition to moving in four directions (as for the coin game), the action space comprises four actions to tag all neighbor agents in the four directions. Moving toward a boundary results in no movement. Only one agent can harvest an apple or tag another agent at a time. The order of actions at each time step is random.