Dynamic Incentivized Cooperation under Changing Rewards

Extended Abstract

Philipp Altmann
LMU Munich
Munich, Germany
philipp.altmann@ifi.lmu.de

Thomy Phan
University of Bayreuth
Bayreuth, Germany
thomy.phan@uni-bayreuth.de

Maximilian Zorn
LMU Munich
Munich, Germany
maximilian.zorn@ifi.lmu.de

Claudia Linnhoff-Popien
LMU Munich
Munich, Germany
linnhoff@ifi.lmu.de

Sven Koenig
University of California & Örebro University
Irvine, USA
sven.koenig@uci.edu

ABSTRACT

Many real-world multi-agent systems are characterized by two simultaneous challenges: strategic tension in social dilemmas and non-stationary reward signals. While *peer incentivization* (PI) has emerged as a decentralized mechanism to promote cooperation in *multi-agent reinforcement learning* (MARL), existing approaches typically rely on fixed or externally scaled incentive magnitudes. When environmental rewards change, due to scaling, shifting, or drift, the relative strength between rewards and incentives can become misaligned, which destabilizes cooperation even when the underlying strategic structure remains unchanged. We analyze this structural sensitivity and argue that reward normalization preserves gradient invariance but does not resolve incentive misalignment in social dilemmas. We then introduce *Dynamic Reward Incentives for Variable Exchange* (DRIVE), a reciprocal shaping mechanism that exchanges reward differences rather than fixed magnitudes. Because these differences are expressed in reward units, they scale proportionally under affine reward changes, preserving the relative influence of environmental rewards and incentives.

Code: https://github.com/philippaltmann/DRIVE

KEYWORDS

Multi-Agent Reinforcement Learning; Emergent Cooperation; Peer Incentivization; Social Dilemmas; Changing Rewards

ACM Reference Format:
Philipp Altmann, Thomy Phan, Maximilian Zorn, Claudia Linnhoff-Popien, and Sven Koenig. 2026. Dynamic Incentivized Cooperation under Changing Rewards: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/TBDR4713

1 MOTIVATION AND BACKGROUND

We can model many real-world AI applications, such as energy management [7], traffic coordination [22], or resource allocation [6], as self-interested, online learning *multi-agent systems (MAS)* where conflicts arise due to opposing goals or shared resources [5]. These interactions are commonly modeled as *social dilemmas* (SDs), where individually rational behavior can lead to collectively suboptimal outcomes [3, 15]. At the same time, real-world reward signals are rarely stationary: task specifications evolve, supply and demand fluctuate, or sensors degrade over time [8, 13]. Even when the interaction's strategic structure remains unchanged, reward magnitudes and offsets may vary. This work considers emergent cooperation at the intersection of SDs and changing rewards.

Setting. We consider a decentralized MARL setting formalized by a Markov game with agents $i \in \mathbb{D}$ selecting actions $a_{t,i}$ at each time step t according to their policy π_i based on their local histories $\tau_{t,i}$ and receiving an individual reward $u_{t,i}$ [16, 25]. Collective performance is measured through social welfare $U = \sum_{i \in \mathbb{D}} \sum_{t=0}^{H-1} u_{t,i}$. To maximize their discounted returns $G_{t,i} = \sum_{k=0}^{\infty} \gamma^k u_{t+k,i}$, we use independent actor-critic learning [4, 24]. The critic estimates the value $\hat{V}_i(\tau_{t,i})$ of the agent's current history $\tau_{t,i}$, and the actor adjusts its policy based on how much the observed outcome deviates from this estimate [12]. This deviation is captured by the *advantage* or *temporal-difference residual* $TD_i(u_{t,i}) = u_{t,i} + \gamma \hat{V}_i(\tau_{t+1,i}) - \hat{V}_i(\tau_{t,i})$, which measures whether the received reward $u_{t,i}$ was better or worse than expected [23]. Although actor-critic methods stabilize learning, independent MARL still optimizes individual returns.

Social dilemmas. In SDs, independently optimizing each $G_{t,i}$ can be misaligned with achieving globally optimal behavior. Specifically we consider 2-player matrix games with the actions C (cooperate) and D (defect), and the payoffs for mutual cooperation R, defection P, exploiting the other T, and being exploited S (cf. Fig. 1a). The *Prisoner's Dilemma* (PD) additionally satisfies $T > R > P > S$, where D is individually rational despite C being socially optimal, as *greed* ($T > R$) and *fear* ($P > S$) often drive agents away from mutual cooperation despite its collective benefit [3, 18, 21].

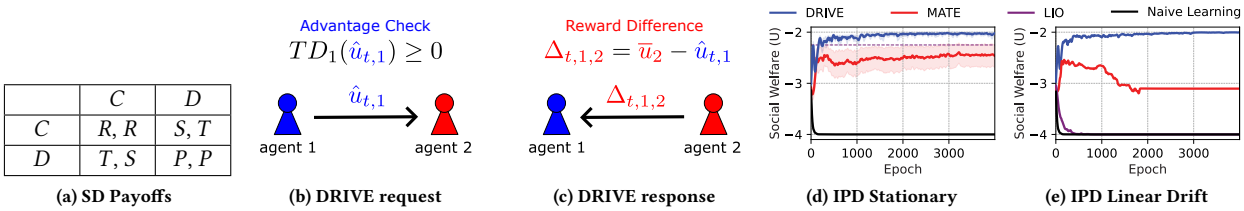
2 MARL UNDER CHANGING REWARDS

Before analyzing incentive mechanisms, it is important to distinguish between changes in *strategic structure* and changes in *reward scale*. Considering a PD under an affine transformation at epoch m:

$$\hat{u}_{t,i} = c_m u_{t,i} + b_m,$$

with $c > 0$, the ordering $\hat{T} > \hat{R} > \hat{P} > \hat{S}$ is preserved. Thus, the strategic structure of the dilemma remains unchanged.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/TBDR4713



(a) SD Payoffs (b) DRIVE request (c) DRIVE response (d) IPD Stationary (e) IPD Linear Drift

Figure 1: (a) Canonical Prisoner's Dilemma payoffs with $T > R > P > S$. We use $T = 0$, $R = -1$, $P = -2$, and $S = -3$. DRIVE exchange protocol: (b) if $TD_i(u_{t,i}) \geq 0$, agent i sends a request; (c) its neighbor responds with the reward difference $\Delta_{t,i,j} = \overline{u}_j - \hat{u}_{t,i}$ using its current average reward $\overline{u}_j$. Average social welfare U in the *Iterated Prisoner's Dilemma* (IPD) for Naive Learning, LIO, MATE, and DRIVE under (d) stationary reward $\hat{u} = u$ and (e) linear drift $\hat{u} = 0.001\, m\, u$. Shaded areas show the 95% confidence intervals. This shows that DRIVE maintains cooperation under drift, while fixed-scale incentive methods degrade as reward scales change.

Naive Learning. In standard policy-gradient MARL without PI, returns are often normalized within each epoch to stabilize learning [26]. Under such normalization, affine transformations of rewards cancel out: scaling and shifting proportionally affect mean and variance, leaving the normalized signal unchanged. However, since the PD inequalities remain unchanged, independent learners still optimize individual returns and converge to defection. Reward normalization preserves learning dynamics, but it does not alter the underlying incentive structure.

Peer Incentivization. PI augments environmental rewards by allowing agents to transfer rewards to one another [10, 20, 28]. In reciprocal or token-based schemes such as MATE [19], a global incentive parameter $x > 0$ determines how strongly cooperation is rewarded or defection is penalized. In a PD, cooperation becomes individually rational only if x exceeds a threshold proportional to payoff gaps such as $(T - R)$ or $(P - S)$. Under affine reward transformations, these gaps behave as follows:

- **Scaling ($c_m \neq 1$, $b_m = 0$).** Payoff differences scale with c_m, while x remains fixed. Thus, even if x was sufficient initially, a large c_m invalidates the cooperation condition. Reward increases weaken fixed incentives; a small c_m can make them overly strong.
- **Shift ($c_m = 1$, $b_m \neq 0$).** A pure shift preserves payoff differences and thus the theoretical threshold. However, if incentives are tuned relative to absolute reward levels or learned value estimates, such shifts can still require retuning in practice [2].
- **Drift (e.g., $c_m = m$).** If reward scales change over epochs, the required threshold evolves accordingly. Any fixed x would require continual adaptation; introducing additional tunable parameters merely shifts the burden of maintaining scale alignment.

These issues also affect learned PI methods, such as *Learning to Incentivize Other learning agents* (LIO) [28], gifting [17], and related incentive methods [9, 11, 14, 27]. Although incentives are learned rather than fixed manually, their magnitudes remain implicitly bounded. When environmental rewards scale or drift, their relative strength to incentives changes unless incentives are proportionally adapted. This failure is therefore structural: gradient normalization preserves update invariance, but not the incentive–reward ratio. As shown in Fig. 1d, LIO and MATE sustain cooperation in the *Iterated Prisoner's Dilemma* (IPD) under stationary rewards but degrade under linear drift (Fig. 1e).

3 DYNAMIC REWARD INCENTIVES

Dynamic Reward Incentives for Variable Exchange (DRIVE) replaces fixed incentive magnitudes with reward-difference exchange [1]. Instead of introducing externally scaled penalties or tokens, agents exchange differences between realized rewards and expected rewards (Fig. 1). Concretely, an agent with non-negative temporal-difference residual issues a request to its peer. The peer responds with a reward difference relative to its own recent average. In this 2-agent scenario, the shaped reward is therefore computed as

$$u_{t,i}^{\text{DRIVE}} = \hat{u}_{t,i} - \left[TD_i(\hat{u}_{t,i}) \geq 0\right] \Delta_{t,i,j} + \left[TD_j(\hat{u}_{t,j}) \geq 0\right] \Delta_{t,j,i}. \quad (1)$$

Intuitively, if an agent benefits from another's action, the gain is reciprocated; if it exploits another, the resulting disadvantage induces a counter-incentive. Because these differences are expressed in the same reward units as $\hat{u}_{t,i}$, they transform proportionally under affine reward changes. Thus, environmental rewards and incentives scale consistently, preserving their relative influence under scaling, shifting, or drift. Fig. 1e, where DRIVE maintains high social welfare despite reward drift, reflects this empirically.

In a Prisoner's Dilemma with $T > R > P > S$, consider unilateral defection (D, C) with payoffs $(\hat{T}, \hat{S})$. If the defector's TD residual is non-negative, it issues a request; the cooperator responds with $\Delta = \bar{u}_2 - \hat{T}$. In steady state $\bar{u}_2 = \hat{S}$, yielding $\Delta = \hat{S} - \hat{T}$. Substituting into Eq. 1 reshapes $(\hat{T}, \hat{S}) \leftrightarrow (\hat{S}, \hat{T})$, while leaving (C, C) and (D, D) unchanged. Unilateral defection is inverted, eliminating incentives for greed and fear and making cooperation a best response.

4 DISCUSSION

Theoretical Perspective. In [1], we prove that reward-difference exchange preserves cooperative equilibria under affine reward transformations. The intuition is that incentives inherit the same transformation structure as environmental rewards.

Empirical Perspective. Furthermore, empirical results in repeated matrix games and sequential social dilemmas with different reward transformations show that DRIVE maintains cooperation under reward drifts, whereas fixed-scale PI methods degrade [1].

Conclusion. MARL with normalization is robust to reward scaling but does not resolve social dilemmas. Static peer incentives solve social dilemmas but are scale-sensitive. DRIVE operates at their intersection, enabling decentralized cooperation under changing rewards.

ACKNOWLEDGMENTS

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. This is a short version of [1].

REFERENCES

[1] Philipp Altmann, Thomy Phan, Maximilian Zorn, Claudia Linnhoff-Popien, and Sven Koenig. 2026. Dynamic Incentivized Cooperation under Changing Rewards. arXiv:2601.06382 [cs.MA] https://arxiv.org/abs/2601.06382

[2] Philipp Altmann, Katharina Winter, Michael Kölle, Maximilian Zorn, and Claudia Linnhoff-Popien. 2025. MEDIATE: Mutually Endorsed Distributed Incentive Acknowledgment Token Exchange. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART*. INSTICC, SciTePress, Porto, Portugal, 33–44. https://doi.org/10.5220/0013091900003890

[3] Robert Axelrod. 1984. *"The Evolution of Cooperation"*. Basic Books, New York. https://doi.org/10.1007/978-3-319-16999-6_1220-1

[4] Daniel S Bernstein, Christopher Amato, Eric A Hansen, and Shlomo Zilberstein. 2009. Policy Iteration for Decentralized Control of Markov Decision Processes. *Journal of Artificial Intelligence Research* 34 (2009), 89–132.

[5] Lucian Buşoniu, Robert Babuška, and Bart De Schutter. 2010. Multi-Agent Reinforcement Learning: An Overview. In *Innovations in Multi-Agent Systems and Applications-1*. Springer, Berlin, Heidelberg, 183–221.

[6] Shuiguang Deng, Zhengzhe Xiang, Peng Zhao, Javid Taheri, Honghao Gao, Jianwei Yin, and Albert Y Zomaya. 2020. Dynamical Resource Allocation in Edge for Trustable Internet-of-Things Systems: A Reinforcement Learning Method. *IEEE Transactions on Industrial Informatics* 16, 9 (2020), 6103–6113.

[7] AL Dimeas and ND Hatziargyriou. 2010. Multi-Agent Reinforcement Learning for Microgrids. In *IEEE PES General Meeting*. IEEE, Minneapolis, Minnesota, USA, 1–8.

[8] Gabriel Dulac-Arnold, Daniel Mankowitz, and Todd Hester. 2019. Challenges of real-world reinforcement learning. *CoRR* abs/1904.12901 (2019), 1–13. arXiv:1904.12901 http://arxiv.org/abs/1904.12901

[9] Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2137–2145.

[10] Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, et al. 2018. Learning with Opponent-Learning Awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Stockholm, Sweden, 122–130.

[11] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. 2018. Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Curran Associates, Inc., Montréal, Canada, 3330–3340.

[12] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. 2019. Human-Level Performance in 3D Multiplayer Games with Population-based Reinforcement Learning. *Science* 364, 6443 (2019), 859–865.

[13] W Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*. Association for Computing Machinery, New York, NY, USA, 9–16.

[14] Michael Kölle, Tim Matheis, Philipp Altmann, and Kyrill Schmid. 2023. Learning to Participate Through Trading of Reward Shares. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART*. INSTICC, SciTePress, Lisbon, Portugal, 355–362. https://doi.org/10.5220/0011781600003393

[15] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, et al. 2017. Multi-Agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems (AAMAS '17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 464–473.

[16] Michael L Littman. 1994. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Machine Learning Proceedings 1994*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 157–163.

[17] Andrei Lupu and Doina Precup. 2020. Gifting in Multi-Agent Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 789–797.

[18] Michael W Macy and Andreas Flache. 2002. Learning Dynamics in Social Dilemmas. *Proceedings of the National Academy of Sciences* 99 (2002), 7229–7236.

[19] Thomy Phan, Felix Sommer, Philipp Altmann, Fabian Ritz, Lenz Belzner, and Claudia Linnhoff-Popien. 2022. Emergent Cooperation from Mutual Acknowledgment Exchange. In *Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)* (Virtual Event, New Zealand). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1047–1055. https://dl.acm.org/doi/abs/10.5555/3535850.3535967

[20] Thomy Phan, Felix Sommer, Fabian Ritz, Philipp Altmann, Jonas Nüßlein, Michael Kölle, Lenz Belzner, and Claudia Linnhoff-Popien. 2024. Emergent Cooperation from Mutual Acknowledgment Exchange in Multi-Agent Reinforcement Learning. *Autonmous Agents and Multi-Agent Systems* 38, 34 (2024), 1–36. https://doi.org/10.1007/s10458-024-09666-5

[21] Anatol Rapoport. 1974. Prisoner's Dilemma — Recollections and Observations. In *Game Theory as a Theory of a Conflict Resolution*. Springer, Heidelberg, Germany, 17–34.

[22] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2016. Safe Multi-Agent Reinforcement Learning for Autonomous Driving. arXiv:1610.03295 [cs.AI]

[23] Richard S Sutton. 1988. Learning to Predict by the Methods of Temporal Differences. *Machine Learning* 3 (1988), 9–44.

[24] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller (Eds.), Vol. 12. MIT Press, Cambridge, MA, USA, 1057–1063.

[25] Ming Tan. 1993. Multi-Agent Reinforcement Learning: Independent versus Cooperative Agents. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 330–337.

[26] Hado P van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. 2016. Learning Values Across Many Orders of Magnitude. *Advances in Neural Information Processing Systems* 29 (2016), 4294–4302.

[27] Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar Duéñez-Guzmán, Alexander Sasha Vezhnevets, and Joel Z Leibo. 2023. A Learning Agent that Acquires Social Norms from Public Sanctions in Decentralized Multi-Agent Settings. *Collective Intelligence* 2, 2 (2023), 14.

[28] Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. 2020. Learning to Incentivize Other Learning Agents. *Advances in Neural Information Processing Systems* 33 (2020), 12.