Towards Socially-Beneficial Multi-Agent Systems: Information, Mechanisms and Dynamics

Doctoral Consortium

Omer Madmon
omermadmon@campus.technion.ac.il
Technion - Israel Institute of Technology
Haifa, Israel

ABSTRACT

Autonomous AI agents operate in economic environments in which outcomes arise from strategic interactions among multiple self-interested decision-makers. Designing such systems to be socially beneficial (in terms of stability, efficiency, or fairness) requires principled models of incentives, information, and learning dynamics. Our research studies these challenges through the lens of algorithmic game theory, focusing on how multi-agent behavior can be steered toward desirable outcomes. Across several interrelated projects, we investigate (i) strategic learning dynamics in competitive content creation and recommendation ecosystems, (ii) information design and persuasion in online platforms, (iii) incentive-compatible mechanisms that promote cooperation among strategic agents, and (iv) language-based economic environments for evaluating LLM-driven agents. Together, this research agenda aims to bridge theory and practice by combining theoretical modeling and analysis with large-scale simulations and empirical research, contributing to the design of socially-beneficial multi-agent systems.

KEYWORDS

Algorithmic Game Theory; Information Design; Mechanism Design; Learning Dynamics; Cooperative AI; Large Language Models

ACM Reference Format:
Omer Madmon. 2026. Towards Socially-Beneficial Multi-Agent Systems: Information, Mechanisms and Dynamics: Doctoral Consortium. 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/GGAC3438

1 INTRODUCTION

Recent advances in generative AI (GenAI) and large language models (LLMs) have enabled the widespread deployment of autonomous agents in real-world economic environments, in which they interact strategically [28]. Understanding and shaping the behavior of AI agents within such multi-agent ecosystems is a central challenge in the study of autonomous agents and multi-agent systems [13, 15, 26, 33, 34, 36]. A core difficulty in these settings is aligning the incentives of self-interested agents with collective objectives, as strategic behavior may lead to undesirable societal outcomes. For instance, competitive content generation can lead to content fluctuations and instability; strategic information disclosure in online markets may distort prices; and cooperation breaks down under communication constraints. Addressing these challenges requires integrating tools from algorithmic game theory [29], information design [7], mechanism design [25], and learning dynamics [8].

Our research develops a unified agenda for designing *socially beneficial multi-agent systems* across a range of application domains, including search and recommendation systems, online retail platforms, data sharing among competitors, and beyond.[1] Our work studies how strategic agents learn and adapt over time, how information disclosure and incentives can be designed to steer behavior toward efficient and fair outcomes, and how cooperation can emerge among self-interested agents despite competition or structural constraints. To bridge the gap between theoretical models and real-world multi-agent systems, we complement theoretical analysis with simulation-based and empirical methodologies. In particular, we develop language-based environments and LLM-driven agent simulations that enable large-scale experimentation in realistic settings. Our goal is twofold: (i) to refine models of strategic behavior in real-world environments, and (ii) to apply and evaluate game-theoretic design principles in real-world multi-agent systems.

2 RESEARCH AGENDA

2.1 Competitive Content Creation Dynamics

In recommendation and search ecosystems, content creators often act strategically to maximize visibility and its associated economic value [16, 27]. Such strategic content manipulation can induce long-term dynamics in which creators fail to converge to stable strategies, leading to persistent fluctuations and ecosystem-level instability. In recent work, we study these dynamics through a game-theoretic lens and design ranking and recommendation mechanisms that guarantee convergence of strategic behavior. We show that deterministic greedy exposure of the most relevant content can lead to instability, and propose an alternative ranking mechanism that induces a *potential game* [21], ensuring convergence of any *better-response dynamics* [18]. In follow-up work, we extend the model to creators who minimize cumulative regret rather than myopically optimizing per-round utility, and use the notion of *socially-concave games* [12] to derive simple sufficient conditions for convergence of any *no-regret dynamics* [17]. Across both projects, we complement the theory with extensive simulations, analyzing

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/GGAC3438

[1] Importantly, the term *socially beneficial* does not assume that modern AI ecosystems are inherently beneficial; many may in fact generate harmful societal dynamics. Instead, we aim to develop formal tools for analyzing and steering multi-agent behavior toward objectives such as welfare, stability, and fairness within these environments.

stability and welfare outcomes for users and publishers as functions of the recommendation mechanism and key ecosystem parameters, such as the number of creators and their integrity costs.

In another project, we study how content creators can be trained to act strategically in competitive search environments [23]. We develop an RL-based framework for fine-tuning *LLM-driven creator agents* and show that learning from repeated ranking competition leads to substantially improved performance over prompting-based approaches [4]. Our results demonstrate that RL provides an effective paradigm for modeling and optimizing strategic content generation in competitive search ecosystems. Building on these foundations, our future work will use LLM-based agents to *simulate content-creator dynamics in realistic textual environments*, systematically comparing their behavior to theoretical predictions.

2.2 Information Design in Online Platforms

In the modern web economy, online platforms mediate interactions between buyers and sellers by *strategically disclosing information*. Such mediation fundamentally shapes incentives, pricing behavior, and welfare outcomes. Relying on economic models of *information design and Bayesian persuasion* [7, 14], we aim to understand how disclosure policies can steer markets toward socially beneficial equilibria. We studied repeated interactions between a platform and a seller, characterizing how the platform can optimally reveal buyer preferences to maximize buyer utility in equilibrium [3]. In a follow-up work, we examined how platforms can reveal buyer information to sellers engaged in *third-degree price discrimination* [6], and proposed a robust and tractable policy that guarantees bounded regret under uncertainty about seller valuations [2].

An important future research direction is to extend information design to collective decision-making environments in which outcomes are determined by *voting rules* [20, 37]. We focus on *supermajority persuasion*, where an informed sender seeks to influence a committee of uninformed voters, and a proposal is adopted only if it receives the support of at least a fixed fraction of voters, rather than a simple majority or unanimity [1, 5, 19]. This setting captures many real-world governance and platform decisions, such as content moderation, standard adoption, and regulatory approval. Our goal is to characterize optimal disclosure policies under supermajority rules, study how persuasion power varies with the voting threshold, and determine how a social planner should choose the supermajority voting rule to maximize social welfare when the sender strategically designs information disclosure to advance objectives that may be misaligned with those of the population.

2.3 Incentive-Compatible Cooperation

Realizing socially beneficial outcomes in multi-agent systems frequently depends on cooperation among self-interested agents [9, 10]. Game theory and mechanism design offer systematic frameworks for engineering incentives that lead to cooperative behavior [25]. Our work seeks to advance new models and mechanisms that foster cooperation in the presence of strategic behavior.

In one line of work, we propose *a strategyproof mechanism for ownership restructuring* that extends the classical BMBY ("Buy Me Out or I'll Buy You Out") mechanism to multi-owner settings [11]. The mechanism enables share buybacks in an incentive-compatible and budget-balanced manner, preserves proportional ownership among remaining shareholders, and allocates control to those who value the asset most, thereby maximizing efficiency subject to proportionality preservation. In a second project, we study *strategic data sharing* between a content creation firm and a competing generative AI platform [35]. Modeling the interaction as a Stackelberg game, we show that equilibrium behavior can involve costly data sharing, where a self-interested firm is willing to pay to share its data. We characterize conditions under which such equilibria constitute Pareto improvements relative to a no-sharing baseline, highlighting how cooperation can arise even between direct competitors. Finally, we extend the *distributed games* framework [22] to environments where communication across locations is constrained by a delay network topology [24]. Focusing on a distributed Prisoner's Dilemma, we derive sufficient conditions under which cooperative equilibria can be sustained despite delayed and limited communication, demonstrating that cooperation remains possible even in severely constrained settings. Future work will further explore incentive-compatible cooperation in complex multi-agent environments, with applications to cybersecurity and agentic AI.

2.4 Language-Based Economic Environments

While classical game-theoretic models provide fundamental insights into the design of socially beneficial multi-agent systems, real-world agentic AI operates in *rich, language-based environments* that are substantially more complex than the stylized settings typically studied in theory. This gap motivates the development of *empirical frameworks* for rigorously evaluating the behavior of AI agents (particularly LLM-driven agents) when they interact with other (human or AI) decision-makers through natural language.

To this end, we developed GLEE, a unified framework and benchmark for simulating *games in language-based economic environments*. GLEE supports canonical economic interactions such as bargaining, negotiation, and persuasion, and enables systematic evaluation of both *individual agent performance* and *collective outcomes*, including efficiency and fairness [32]. Using GLEE, we collected and analyzed large-scale LLM vs. LLM and human vs. LLM interaction data. Our analysis revealed non-trivial patterns in strategic behavior, outcome distributions, and the relationship between language, information, and communication structure. By standardizing environments, metrics, and experimental protocols, GLEE facilitates controlled comparisons across agents and economic settings.

A complementary line of work focuses on *human choice prediction* in strategic, language-based interactions. Using a custom-developed mobile game, we collected large-scale data on repeated persuasion dynamics and studied ML approaches for predicting human decisions. This includes simulation-based off-policy data generation [30] and LLM-driven data synthesis [31], both of which substantially improve predictive accuracy and shed light on the behavioral mechanisms underlying human decision-making.

Building on these foundations, our future work will move beyond evaluation toward *steering agent behavior* in language-based economic settings. In particular, we aim to develop methods that allow LLM-based agents to operate along a controllable spectrum between *human-like behavior* and *fully rational decision-making*, depending on the application.

REFERENCES

[1] Ricardo Alonso and Odilon Câmara. 2016. Persuading voters. *American Economic Review* 106, 11 (2016), 3590–3605.

[2] Itai Arieli, Yakov Babichenko, Omer Madmon, and Moshe Tennenholtz. 2025. Robust price discrimination. *Games and Economic Behavior* 154 (2025), 377–395.

[3] Itai Arieli, Omer Madmon, and Moshe Tennenholtz. 2024. Reputation-based persuasion platforms. *Games and Economic Behavior* 147 (2024), 128–147.

[4] Niv Bardas, Tommy Mordo, Oren Kurland, and Moshe Tennenholtz. 2025. Automatic Document Editing for Improved Ranking. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 2779–2783. https://doi.org/10.1145/3726302.3730168

[5] Arjada Bardhi and Yingni Guo. 2018. Modes of persuasion toward unanimous consent. *Theoretical Economics* 13, 3 (2018), 1111–1149.

[6] Dirk Bergemann, Benjamin Brooks, and Stephen Morris. 2015. The limits of price discrimination. *American Economic Review* 105, 3 (2015), 921–957.

[7] Dirk Bergemann and Stephen Morris. 2019. Information design: A unified perspective. *Journal of Economic Literature* 57, 1 (2019), 44–95.

[8] Ronen Brafman and Moshe Tennenholtz. 2002. Efficient learning equilibrium. *Advances in Neural Information Processing Systems* 15 (2002).

[9] Vincent Conitzer and Caspar Oesterheld. 2023. Foundations of cooperative AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 15359–15367.

[10] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630* (2020).

[11] Gal Danino, Moran Koren, and Omer Madmon. 2026. The Multi-BMBY Mechanism: Proportionality-Preserving and Strategyproof Ownership Restructuring in Private Companies. *Journal of Economics & Management Strategy* 35, 1 (2026), 50–58.

[12] Eyal Even-Dar, Yishay Mansour, and Uri Nadav. 2009. On the convergence of regret minimization dynamics in concave games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 523–532.

[13] Michael N Huhns and Larry M Stephens. 1999. Multiagent systems and societies of agents. *Multiagent systems: a modern approach to distributed artificial intelligence* 1 (1999), 79–114.

[14] Emir Kamenica and Matthew Gentzkow. 2011. Bayesian persuasion. *American Economic Review* 101, 6 (2011), 2590–2615.

[15] Steffi Knorn, Zhiyong Chen, and Richard H Middleton. 2015. Overview: Collective control of multiagent systems. *IEEE Transactions on Control of Network Systems* 3, 4 (2015), 334–347.

[16] Oren Kurland and Moshe Tennenholtz. 2022. Competitive search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2838–2849.

[17] Omer Madmon, Idan Pipano, Itamar Reinman, and Moshe Tennenholtz. 2025. On the Convergence of No-Regret Dynamics in Information Retrieval Games with Proportional Ranking Functions. In *The Thirteenth International Conference on Learning Representations*.

[18] Omer Madmon, Idan Pipano, Itamar Reinman, and Moshe Tennenholtz. 2025. The search for stability: Learning dynamics of strategic publishers with initial documents. *Journal of Artificial Intelligence Research* 83 (2025).

[19] Anthony J McGann. 2004. The tyranny of the supermajority: How majority rule protects minorities. *Journal of Theoretical Politics* 16, 1 (2004), 53–77.

[20] Reshef Meir. 2018. *Strategic voting*. Morgan & Claypool Publishers.

[21] Dov Monderer and Lloyd S Shapley. 1996. Potential games. *Games and economic behavior* 14, 1 (1996), 124–143.

[22] Dov Monderer and Moshe Tennenholtz. 1999. Distributed games. *Games and Economic Behavior* 28, 1 (1999), 55–72.

[23] Tommy Mordo, Sagie Dekel, Omer Madmon, Moshe Tennenholtz, and Oren Kurland. 2025. RLRF: Competitive Search Agent Design via Reinforcement Learning from Ranker Feedback. *arXiv preprint arXiv:2510.04096* (2025).

[24] Tommy Mordo, Omer Madmon, and Moshe Tennenholtz. 2025. Cooperation Under Network-Constrained Communication. *arXiv preprint arXiv:2511.05290* (2025).

[25] Roger B Myerson. 2008. Mechanism design. In *The New Palgrave Dictionary of Economics*. Springer, 1–13.

[26] David C Parkes and Michael P Wellman. 2015. Economic reasoning and artificial intelligence. *Science* 349, 6245 (2015), 267–272.

[27] Kun Qian and Sanjay Jain. 2024. Digital content creation: An analysis of the impact of recommendation systems. *Management Science* (2024).

[28] David M Rothschild, Markus Mobius, Jake M Hofman, Eleanor W Dillon, Daniel G Goldstein, Nicole Immorlica, Sonia Jaffe, Brendan Lucier, Aleksandrs Slivkins, and Matthew Vogel. 2025. The Agentic Economy. *arXiv preprint arXiv:2505.15799* (2025).

[29] Tim Roughgarden. 2010. Algorithmic game theory. *Commun. ACM* 53, 7 (2010), 78–86.

[30] Eilam Shapira, Omer Madmon, Reut Apel, Moshe Tennenholtz, and Roi Reichart. 2025. Human Choice Prediction in Language-based Persuasion Games: Simulation-based Off-Policy Evaluation. *Transactions of the Association for Computational Linguistics* 13 (2025), 980–1006.

[31] Eilam Shapira, Omer Madmon, Roi Reichart, and Moshe Tennenholtz. 2024. Can llms replace economic choice prediction labs? the case of language-based persuasion games. *arXiv preprint arXiv:2401.17435*.

[32] Eilam Shapira, Omer Madmon, Itamar Reinman, Samuel Joseph Amouyal, Roi Reichart, and Moshe Tennenholtz. 2024. Glee: A unified framework and benchmark for language-based economic environments. *arXiv preprint arXiv:2410.05254*.

[33] Yoav Shoham and Kevin Leyton-Brown. 2008. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.

[34] Peter Stone and Manuela Veloso. 2000. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots* 8, 3 (2000), 345–383.

[35] Boaz Taitler, Omer Madmon, Moshe Tennenholtz, and Omer Ben-Porat. 2025. Data Sharing with a Generative AI Competitor. *arXiv preprint arXiv:2505.12386* (2025).

[36] Michael Wooldridge. 2009. *An introduction to multiagent systems*. John wiley & sons.

[37] Peyton Young. 1995. Optimal voting rules. *Journal of Economic Perspectives* 9, 1 (1995), 51–64.