Proceedings of the Third International Workshop on Citizen-Centric Multiagent Systems 2025 (C-MAS 2025)

Co-located with the International Conference on Autonomous Agents and Multiagent Systems (AAMAS'25)

Detroit, Michigan, USA

Yali Du[1], Nadin Kokciyan[2], Behrad Koohy[3], Fernando P. Santos[4],
Sebastian Stein[3], and Vahid Yazdanpanah[3]

[1]King's College London
[2]University of Edinburgh
[3]University of Southampton
[4]University of Amsterdam

20 May 2025

Welcome to the third edition of C-MAS, the International Workshop on Citizen-Centric Multiagent Systems. Over the past two years, C-MAS has focused on reshaping how we think about AI and multiagent systems in relation to society and role of end users. We continue to challenge the conventional view of users as passive data sources or service consumers. Instead, we emphasise the role of citizens as active agents with their own goals, preferences, and responsibilities within sociotechnical systems. As AI technologies increasingly shape our public spaces, communities, and infrastructures, a citizen-centric perspective become crucial for ensuring these systems are inclusive, trustworthy, and socially beneficial. C-MAS 2025 builds on the foundations laid in 2023 and 2024, pushing further into questions of participation, agency, and impact.

This year, we are expanding our focus to highlight not only the design of citizen-centric MAS but also their deployment and evaluation in real-world contexts. Our sessions cover key themes such as "Empowering Citizens in Critical Services", "Modelling Human Needs in Shared Environments", "Multiagent Learning for Public Decision-Making", and "Ethics, Fairness, and Normative Reasoning". Through these discussions, we aim to bridge the gap between research prototypes and impactful, citizen-focused AI solutions. We are excited to welcome an interdisciplinary community of researchers all contributing to a shared vision: AI systems that serve and empower the citizens.

Further details about CMAS-2025 are available at: https://sites.google.com/view/cmas25

Contents

1 Keynote: Tackling Societal Challenges with Multi-Agent Systems: Bridging Theory and Practice

Keynote by Dr. Fei Fang, Carnegie Mellon University

This talk will cover real-world challenges—from wildlife conservation to public health—often involve coordination or competition among multiple agents. In this talk, Prof. Fei Fang will present their work on multi-agent systems and their applications in critical societal domains, including anti-poaching efforts, food rescue operations, and mental health support. She will discuss key theoretical advancements, practical deployments, and the lessons learned in translating multi-agent research into real-world impact.

Dr. Fei Fang is an Associate Professor at the Software and Societal Systems Department in the School of Computer Science at Carnegie Mellon University. Before joining CMU, she was a Postdoctoral Fellow at the Center for Research on Computation and Society (CRCS) at Harvard University, hosted by David Parkes and Barbara Grosz. She received her Ph.D. from the Department of Computer Science at the University of Southern California advised by Milind Tambe (now at Harvard). Fei's research lies in the field of artificial intelligence and multi-agent systems, focusing on integrating machine learning with game theory. Her work has been motivated by and applied to security, sustainability, and mobility domains, contributing to the theme of AI for Social Good. She is the recipient of the Allen Newell Award for Research Excellence 2023, 2022 Sloan Research Fellowship, and IJCAI-21 Computers and Thought Award. She was named to IEEE Intelligent Systems' "AI's 10 to Watch" list for 2020. Her work has won the Best Paper Award at GameSec'23, Deployed Application Award at IAAI'23, Best Paper Honorable Mention at HCOMP'22, Best Paper Runner-Up at AAAI'21, Distinguished Paper at IJCAI-ECAI'18, Innovative Application Award at IAAI'16, the Outstanding Paper Award in Computational Sustainability Track at IJCAI'15. She received an NSF CAREER Award in 2021. Her dissertation is selected as the runner-up for IFAAMAS-16 Victor Lesser Distinguished Dissertation Award, and is selected to be the winner of the William F. Ballhaus, Jr. Prize for Excellence in Graduate Engineering Research as well as the Best Dissertation Award in Computer Science at the University of Southern California.

2 Empowering Citizens in Critical Services

2.1 MEAL: Model of Empathy Augmented Logistics for Food Security

Meal: Model of Empathy Augmented Logistics for Food Security

Seoyeong Park[1][0009−0001−4402−6160] and Munindar P. Singh[1][0000−0003−3599−3893]

North Carolina State University, Raleigh, NC, USA
spark43@ncsu.edu, mpsingh@ncsu.edu

Abstract. Millions globally lack access to nutritious food, experiencing food insecurity. Efforts to address food insecurity seek to provide consumers food that may be *rescued* (i.e., what warehouses or grocers would otherwise soon discard as unusable), directly donated, or acquired using governmental funds.

Current approaches produce allocations that optimize global objectives to store and move food efficiently across food banks. However, they largely overlook consumer preferences and constraints. As a result, the resulting allocations lead to consumers either using foods they do not care for or discarding such foods, leading to food waste.

This paper presents a new model, studied via human study and agent-based simulation, that shows how incorporating the consumer perspective on par with the provider perspective can lead to better outcomes overall. We find that persuasive messages that include individual circumstances and the social context can promote prosociality and empathy.

Keywords: Food security · Multiagent system · Agent-based simulation

1 Introduction

Food insecurity is the condition of a household having poor access to adequate food and reduced quality of food intake [8]. One-eighth (approximately 17 million) of US households experience food insecurity [8], and it is a critical global concern [7].

The US food bank system is a nonprofit organization that reduces food waste and alleviates food insecurity by collecting, storing, and distributing food to those in need [1]. The federal government provides funding and capabilities to procure, store, transport, and distribute food [8]. Local food banks (providers) may receive donations from organizations, retailers, and individuals as well as allocations from regional food banks. Volunteers sort and distribute food to consumers and sometimes to smaller sites called food pantries. A consumer is a household experiencing food insecurity. Consumers deserve not only to satisfy their health-based or cultural dietary needs and to have a choice on what they eat, albeit limited by what is available.

Ensuring equitable distribution is difficult when supplies are in short supply, and preferences are diverse. Thus, a traditional approach may end up giving its limited supply of milk to a household without children while a household with children has to do without. Or, it might allocate starchy foods to a person with diabetes. Current research addresses logistic efficiency [2, 12] or concentrating on consumers' tastes [11], but not on both aspects together.

We propose **Meal** for Model of Empathy Augmented Logistics for Food Security. Meal allocates food by considering both consumer needs and societal objectives such as reducing food waste and improving equity. Meal's **novelty** thus lies in combining prosociality with a multistakeholder model of food security. Through extensive simulation experiments, we find that Meal reduces waste and increases satisfaction in distributing food items compared to models that consider only one side, either consumers or providers. Through a *human study*, we find that persuasive messages, especially those that fit individual circumstances and the social context, can promote prosociality.

2 Motivation for Meal

In an ideal world, everyone would get the food items they most prefer. However, it is impossible to match everyone's preferences with constraints. Previous approaches to promoting food security through sharing food with those in need have generally taken a rigid stance. In these approaches, an organization such as a food bank, which has all the power and the food, decides how to allocate it to food-insecure households. Besides the obvious challenges of not accommodating the wishes of the intended recipients of the food, this approach leads to greater food waste system-wide because foods that do not match the constraints and preferences of the recipients cannot be used by them. This top-down allocation inevitably ends up with some consumers not receiving their preferred items, which not only leaves them less satisfied but also worsens food waste. Therefore, we consider restructuring the problem such that other acceptable allocations can be found. Our approach builds on key principles: social welfare, equity, prosociality, and empathy.

2.1 Stakeholders

We consider two main types of stakeholders. *Consumers* are households served through our recommendation system. They aim to acquire food items that align with their preferences and needs. This consumer-centric perspective emphasizes the importance of enhancing consumer satisfaction and personalized experiences for food allocation [3]. As consumers interact with the system, their preferences for food items are constantly captured and refined. These preferences evolve over time and are shaped by factors such as age, health status, dietary constraints, household status, and willingness to make prosocial choices [4, 5]. The agent learns these dynamics by reflecting consumer feedback toward recommended food items. This learning process allows the agent to provide recommendations matching a consumer's tastes and current needs.

Providers seek to improve the effective distribution of the available food. This entails reducing food waste, maximizing the distribution of food, and meeting the needs of their community while providing food items that suit consumer preferences. The provider prioritizes not merely using in-stock items but also fulfilling consumer requests as closely as possible [3]. However, they might propose less-preferred alternatives when necessary. The provider intends to trigger empathy and gently nudge consumers to accept alternatives through social and psychological factors that influence decision-making.

2.2 Research Questions

Accordingly, this study investigates these research questions.

$\mathbf{RQ}_{\mathbf{prosociality}}$ How can MEAL produce equitable allocations by incorporating a dynamic multistakeholder context (consumers and providers) and supporting prosocial behavior among consumers?

$\mathbf{RQ}_{\mathbf{persuasion}}$ Do persuasion and empathy influence human decisions about food and prosocial behavior?

3 Empirical Evaluation with Humans

Even if MEAL recommends substitutes that are mostly consistent with preferences, simply offering those without any context or with a generic explanation is less effective and unhelpful for consumers. To validate our assumptions on human behavior and prosociality underlying our simulation, we conducted an IRB-approved human study on consumer decision-making. Our study shows personalized, context-rich persuasive messages may improve engagement compared to simple and generic ones.

We observed no significant difference in decision-making with or without a persuasive message. The acceptance rates were similar for *No persuasion* and *Persuasion*, 62.5% in the former and 63.7% in the latter. Applying the two-proportion Z-test [10] produced a p-value of 0.7, indicating no significant difference. This indicates that the consumers are highly likely not affected by persuasive situations when the system provides justification and context, implying that the persuasive message used in the study was too weak or generic to resonate with the participants' priorities.

Similarly, we found no statistically significant difference in consumer satisfaction: the mean of 3.57 *No persuasion* and the mean of 3.43 *Persuasion*, with a Mann-Whitney U test [6] p-value of 0.193. The results show that consumer satisfaction was not greatly affected by the given persuasive message. This indicates that the observed increase in acceptance rate with persuasion may not necessarily translate to a corresponding increase in consumer satisfaction. In other words, simply encouraging to accept substitutes may not be accompanied to enhance the consumer's experience.

Understanding what motivates consumers to accept recommendations is crucial. The survey given at the end of the study revealed that the participants are most likely to accept if the alternatives are what they like or similar to their original choices in terms of taste, type of food, or nutritional value; in other words, familiarity matters.

Consumers may measure their satisfaction not only with fulfilling personal desires but also by feeling rewarded for helping others. Almost all survey respondents answered that they would highly likely change their decision of refusing a recommended item regardless of their personal situations if they know their choice helps promote social well-being, unless they have strong dietary restrictions.

4 Model Design

Our goal is to simultaneously maximize consumer satisfaction and maximize the provider's benefit. The agent understands stakeholders' values, the future state of the world for each action it can perform, and the social experience its consumer will derive for each action it can perform. Then, since we cannot maximize both objectives, the agent moderates to achieve an optimal trade-off between two stakeholders.

We now formalize our problem setup. We have a set of consumers U and a set of food items F, where each consumer in U has profile information and unique food preferences toward each food item in F. Each item in F carries attributes that reflect its importance in consumption priority and benefits to the provider. These attributes include multiple factors, such as inventory capacity, expiration date, and perishability, shaping the provider benefits $c_{u,d,t}$ associated with each recommendation happening at time step t. Within this dynamic framework, $d_{u,f,t} \in D$ represents a recommendation for consumer u at a specific time step t. It contains two attributes: a recommended food item and a binary indicator of whether it is accepted. Subsequently, we define that consumer satisfaction $h_{u,d,t}$ comes as ratings at a time step t, ranging from 0 (no preference or experience) to 5 (extremely like). The provider's benefit c is determined by the aggregate score of accepted food items, scaling to the same range as h. These scores are updated in real time as allocations are made.

The problem involves finding the optimal way to distribute the available food to consumers over time while considering their preferences and impact on the community, in other words, managing the trade-off between these two objectives. To balance these objectives, a weighted sum of consumer satisfaction H and provider benefit C is used with a weighting factor denoted as ω ($0 \leq \omega \leq 1$). We choose the optimal value of ω that maximizes both H and C. Therefore, the agent's overall reward for the decision-making objective is a weighted combination of satisfaction and provider benefit.

By using Q-learning [9], our model effectively adapts to dynamic changes in consumers' needs, food availability, and other factors and incorporates long-term interaction into their decision-making process.

5 Results

Our study considers three baselines: random recommendation, consumer-focused, and provider-focused approaches.

Random recommendation Recommends items randomly from in-stock inventory, regardless of consumer preferences or provider benefit. This baseline disregards fairness and trust.

Consumer-focused Solely prioritizes consumers' preferences based on their past interactions and preferences. This model is equivalent to assigning a weighting factor ω of 1 and completely ignores provider benefit.

Provider-focused Solely prioritizes the provider-side operation exclusively and disregards consumer preferences. It is equivalent to assigning a weighting factor ω of 0.

5.1 Consumer Satisfaction

Consumers find greater satisfaction with recommendations that consider both consumer preference and society's welfare. This trade-off indicates that Meal fulfills the intended objectives even though it might sacrifice some provider benefits.

The provider-focused model delivers the highest cumulative provider benefit, and the consumer-focused model achieves the lowest provider benefit. The provider benefit decreases as the weight assigned to the provider decreases, in other words, it increases inversely related to ω. Consumer satisfaction visibly improves, unlike what we originally expected both stakeholders to sacrifice to some extent if we set a parameter for the reward. The evenly considered ($\omega = 0.5$) model and the optimal ($\omega = 0.2$) model outperform the consumer-focused model in terms of getting higher consumer satisfaction. It indicates that Meal recommends items that consumers like more.

This implies that the weighted models distribute resources in a way that actually benefits both consumers and providers more. By incorporating the provider's perspective, Meal achieves a more efficient and equitable allocation, meaning that a greater number of consumers are served or a greater number of consumers get better at matching their preferred items among the available inventory.

5.2 Acceptance of Recommendations

How much the model skews to consumer satisfaction affects the acceptance rate. The higher the weight on consumer preferences, the higher the acceptance rate. The gap in the acceptance rate between the consumer-focused and provider-focused models differs notably. The consumer-focused model dominates all other models, particularly the provider-focused and random recommendation. We could observe that the acceptance rate gradually drops in the provider-focused model, unlike increasing in other models. This result implies that when the provider recommends items that need to be sold quickly, without paying much attention to whether they match the consumer's preferences, consumers often find these recommendations less appealing. As a result, they are more likely to reject them.

Interestingly, models that incorporate some preference weighting tend to converge to acceptance rates that are similar to the consumer-focused model, with only slight differences of less than 1%. This observation indicates that while the consumer-focused model has the strongest alignment with consumer preferences and needs, weighted models still achieve comparable acceptance rates. It means that consumers are highly likely to accept substitutions even when recommendations are not perfectly tailored but reasonably close to their preferences, which eventually results in a better overall resource allocation.

5.3 Potential in Food Waste Reduction

Our result represents the estimated percentage of food wasted at each timestep. Waste after acceptance is excluded but all other expired food items are included. It shows that the percentage of food waste increases early stages but gently decreases after a certain point. The optimal model ($\omega = 0.2$) lowers the waste below the consumer-focused model and is close to the provider-focused model. That is, the optimal model shows only a small difference in food waste compared to the provider-focused model, even though the model considers the provider's benefit less.

6 Limitations and Future Work

Our proposed model faces some limitations. First, Meal elides nutritional factors and health considerations and recommends items solely relying on explicit preferences toward each food item given by consumers. Likewise, attributes such as socioeconomic background, culture, religion, and other diversity across communities remain challenging for optimization.

Incorporating additional stakeholder types would provide a more holistic view but complicate ensuring well-being, fairness, and trust among the stakeholders.

7 Conclusion

Achieving equitable food distribution requires a multifaceted endeavor that meets various goals. Meal seeks to optimize the allocation strategy toward maximizing the rewards for consumer satisfaction and provider benefit, employing Q-learning. Our findings highlight that the right balance of the stakeholders' objectives enhances consumer satisfaction while maximizing provider benefits. Our experiments simulate the society aligning with theoretical literature and other empirical findings in the relevant fields. Such alignment reinforces the robustness and applicability of our proposed method in real-world scenarios.

Acknowledgments. This work is supported by the NSF under grant 2125600.

References

1. America, F.: Our work (2024), https://www.feedingamerica.org/our-work, [Accessed 2024-10-06]
2. Iiyama, T., Kitakoshi, D., Suzuki, M.: Optimizing food allocation in food banks with multi-agent deep reinforcement learning. In: Proceedings of the International Conference on Technologies and Applications of Artificial Intelligence (TAAI). pp. 203–208. IEEE, Tainan, Taiwan (Dec 2022). https://doi.org/10.1109/TAAI57707.2022.00045, https://doi.org/10.1109/TAAI57707.2022.00045
3. Miller, M., Holston, D., Losavio, R.: Client-choice food pantry guide (2021), https://www.lsuagcenter.com/profiles/aiverson/articles/page1614284604730, [Accessed 2024-10-10]
4. National Academies of Sciences, Engineering, and Medicine: Understanding food waste, consumers, and the US food environment. In: Schneeman, B.O., Oria, M. (eds.) A National Strategy to Reduce Food Waste at the Consumer Level. The National Academies Press, Washington, DC (2020). https://doi.org/10.17226/25876, https://nap.nationalacademies.org/catalog/25876/a-national-strategy-to-reduce-food-waste-at-the-consumer-level
5. Ogundijo, D.A., Tas, A.A., Onarinde, B.A.: Age, an important sociodemographic determinant of factors influencing consumers' food choices and purchasing habits: An english university setting. Frontiers in Nutrition **9**, 858593 (2022). https://doi.org/10.3389/fnut.2022.858593, https://www.frontiersin.org/journals/nutrition/articles/10.3389/fnut.2022.858593
6. Sprent, P., Smeeton, N.C.: Methods for Two Independent Samples, chap. 6, pp. 151–161. CRC press, New York (2016)
7. Tahtinen, L., Costa, N., Long, Y., Roberts-Harry, G.: SDG Good Practices: A Compilation of Success Stories and Lessons Learned in SDG Implementation. United Nations Department of Economic and Social Affairs, New York, 2 edn. (2022)
8. USDA: Food security and nutrition assistance (2023), https://www.ers.usda.gov/data-products/ag-and-food-statistics-charting-the-essentials/food-security-and-nutrition-assistance/, [Accessed 2024-10-06]
9. Watkins, C.J., Dayan, P.: Q-learning. Machine Learning **8**, 279–292 (1992). https://doi.org/10.1007/BF00992698, https://doi.org/10.1007/BF00992698
10. Webb, R.: Two Proportion Z-Test and Confidence Interval, chap. 9.3, p. 320. Portland State University Library, Portland, Oregon (2021), https://pdxscholar.library.pdx.edu/pdxopen/36/
11. Yang, L., Hsieh, C.K., Yang, H., Pollak, J.P., Dell, N., Belongie, S., Cole, C., Estrin, D.: Yum-Me: A personalized nutrient-based meal recommender system. Transactions on Information Systems (TOIS) **36**(1), 1–31 (2017). https://doi.org/10.1145/3072614, https://doi.org/10.1145/3072614
12. Zoha, N., Hasnain, T., Ivy, J.: Tradeoff between geographic and demographic equity in food bank operations. In: IISE Annual Conference Proceedings. pp. 1–6. Institute of Industrial and Systems Engineers (IISE), Seattle, Washington (may 2022), https://www.proquest.com/scholarly-journals/tradeoff-between-geographic-demographic-equity/docview/2715836589/se-2

Appendix

A.1 US Food Bank System

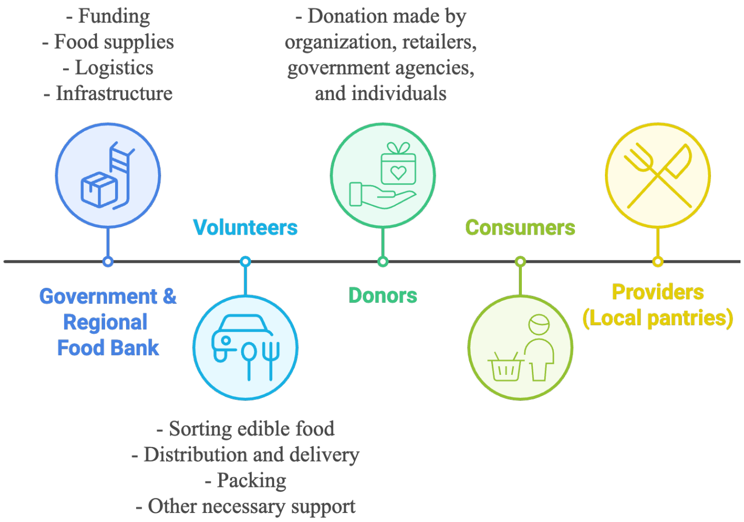

- Funding
- Food supplies
- Logistics
- Infrastructure
- Donation made by organization, retailers, government agencies, and individuals
Volunteers
Consumers
Government & Regional Food Bank
Donors
Providers (Local pantries)
- Sorting edible food
- Distribution and delivery
- Packing
- Other necessary support

Fig. 1: Food distribution system, based on the US setting.

A.2 Concept of Operations in Meal

We envision that consumers register with the food-sharing app by providing their profiles (e.g., household information). Consumers indicate preferences for some food items, e.g., fresh fruits and vegetables, milk, and whole grains. Then, they request food items as they need. Based on the inventory availability, community demands, and the consumer's profile and past selections, the app recommends alternative items from the same categories if one or more requested items are not available. The consumer can choose to accept or reject these substitutions and indicate their satisfaction with the accepted items, which the app uses to refine its suggestions.

Fig. 2 illustrates our conception. The agent serves as a mediator between consumers and a provider using consumer preferences and profiles to form the foundation for personalized recommendations. The agent received the provider's inventory information to make accurate up-to-date recommendations. Then, it aggregates demand and trends, estimates the level of prosociality of consumers and the goodness of food items, and processes interactions so that all parties benefit.

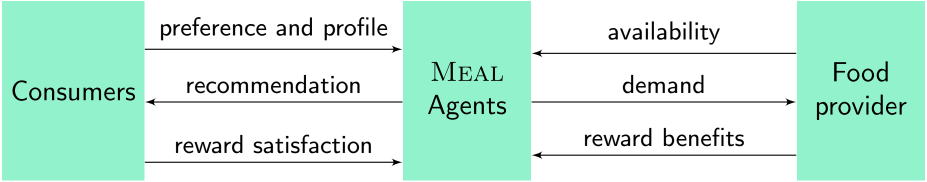


Fig. 2: Model architecture.

In general, the app cannot always recommend the consumer's most preferred items. For instance, if apples have a higher demand than available stock, the app might suggest oranges. Doing so helps ensure as many people as possible get what they need and keeps the food bank running smoothly. Thus, consumers and providers have different perspectives. MEAL recognizes complexity by modeling consumers focusing on household needs and preferences, and a provider managing availability and community demand.

A.2.1 Model Formulation

Formally, we define the above problem as a Partially Observable Markov Decision Process (POMDP) where an agent (recommender) interacts with the environments (consumers and food provider) over time to maximize cumulative rewards of combined benefits. $\langle S, A, T, R, O, \Omega, \gamma \rangle$, where $s \in S$ is a finite set of states (i.e., consumer preferences and profiles, inventory status), $a \in A$ is a finite set of actions (i.e., the possible recommendations), T is a set of transition probabilities between states (i.e., the probability of acceptance), O is a set of observations (i.e., whether the recommendation is taken or not, consumers' satisfaction feedback), Ω is a set of conditional observation probabilities of receiving an observation $o \in O$ after taking action $a \in A$ at state s, R is a reward function (i.e., a combination of consumer satisfaction and provider's benefit controlled by the weighting factor ω, as defined in Equation 2), and $\gamma \in [0, 1)$ is the discount factor.

$$r_\omega = \omega \cdot h + (1 - \omega) \cdot c \tag{1}$$

$$\omega^* = \arg \max_{\omega \in [0,1]} r(\omega) \tag{2}$$

A.3 Empirical Study Design

To conduct this study, we built a simple app that follows the streamlined flow of food requests and recommends replacements. We recruited 49 (adult, US-based) volunteers without any restrictions to ensure diverse representation.

The study involves two sessions of three food-requesting flows each. One session does not have persuasive messages when recommending replacements; the

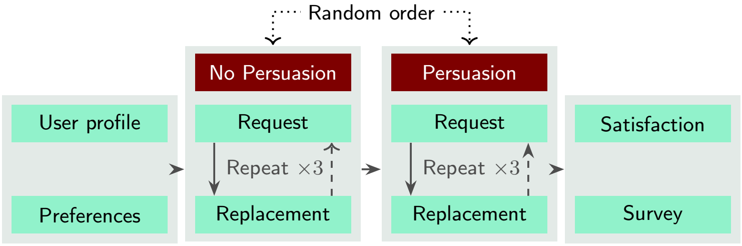


Fig. 3: User study design.

other does. All 49 participants completed both the *Persuasion* and *No persuasion* sessions but in randomized order to consider potential dropout in the middle of the study. In each episode, the participants choose items from a list of fruits, vegetables, and meats. In the treatment, we replace two items in each food category after each request, and the participants can choose to accept or reject the replacements. At the end of the sessions, the participants rated how satisfied they were with the replacements they accepted.

Table 1: Data summary and results

	No persuasion	**Persuasion**
Total responses	515	463
Accepted	322	295
Rejected	193	168
No satisfaction response	91	54
Acceptance percentage	62.52	63.71
Mean satisfaction	3.57	3.43
Median satisfaction	4	4

A.4 Experimental Setting

We evaluate our model through simulations to understand how prosocial decisions are made throughout interactions. The simulated environment comprises data consisting of three sets: consumer profiles, preference ratings, and food inventory. Since it is hard to acquire real-world food preference data and food bank availability, we arbitrarily approximated the values of food items in our simulation by seeding the survey results of food pantry needs [1].

A.4.1 Consumer Profile and Prosociality

The main agents in our model are the consumers. We have crafted a consumer community with unique profiles. For simplicity, each consumer's profile includes age, whether they have one or more children, whether they have dietary restrictions or disease, family size, and ratings towards food items. We set 33% of consumers as aged over 65 and 45% of consumers as having a child. The family size distribution followed the statistics derived from a survey: the mean is three, and the standard deviation is two [1].

A consumer may accept or reject a recommendation. The probability of acceptance hinges on two factors: the consumer's preference and inherent willingness to yield. Consumers don't know how much the provider gains from their decisions. Ratings for particular items may be undefined. If undefined, we estimate satisfaction with the most similar consumer preferences using cosine similarity.

A.4.2 Food Inventory

Our simulation necessitates a comprehensive and realistic dataset that encompasses not just the items but also their attributes. We obtained a food list from [2] (169 different items) and classified it into six categories that people request every day, which are meat, fruits and vegetables, dairy, eggs, cooking items (like oils and seasoning), and others. However, since the [2] data lacks the specific attributes we need, we augmented attributes with feasible assumptions as close to demands mentioned in [1]. For simplicity, we limit to considering quantity, expiration date, and perishability as key components of setting urgency of allocation.

A.4.3 Trade-Offs: Provider versus Consumer

We evaluate various weightings to determine the optimal value of ω, as in Equation 2. We observe that the weighted models surpass the consumer-focused model in cumulative satisfaction, demonstrating the effectiveness of Meal.

A.5 Visualizations of Results

To verify our model, we conduct simulations with 1,000 agents, each corresponding to consumers, one agent corresponding to the provider, and Meal agent acting as a moderator between the consumers and the provider.

The parameter ω ranges between completely provider-focused valuation ($\omega = 0$) and completely consumer-focused ($\omega = 1$), with increment of 0.1. H and C are updated each time a particular recommendation is taken.

To evaluate our model's performance, we consider two distinct values for the weighing factor (ω): 0.2, optimal in our setting determined by Equation 2, and 0.5, which evenly considers both sides. The results consistently show that our model with the optimal value of the weighting factor achieves our goal of satisfying both stakeholders' objectives. The model is trained with a learning rate (α) of 0.1, a discount factor (γ) of 0.9, an exploration rate (ϵ) of 0.1, and a prosociality weight (β) of 0.1.

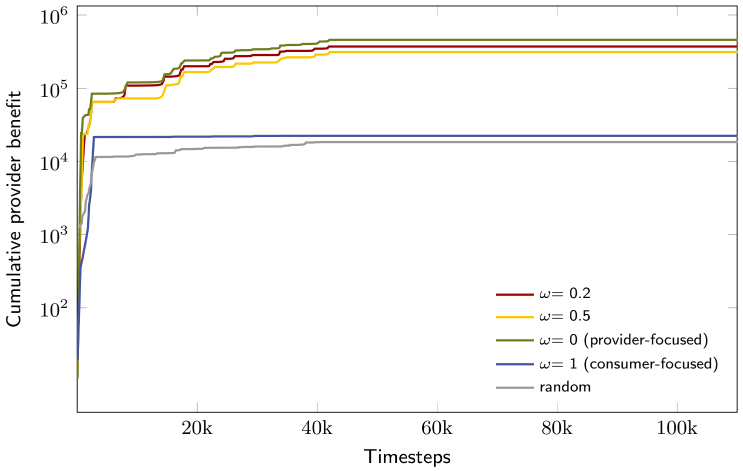


Fig. 4: Cumulative provider benefit. The provider-focused model gains the most while the consumer-focused model gains the least.

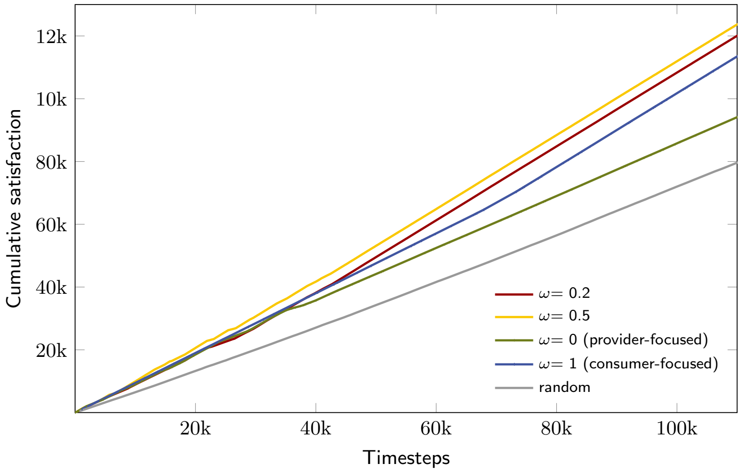


Fig. 5: Cumulative consumer satisfaction. Weighted models have the potential to achieve higher satisfaction.

References

1. Caspi, C.E., Davey, C., Barsness, C.B., Gordon, N., Bohen, L., Canterbury, M., Peterson, H., Pratt, R.: Needs and preferences among food pantry clients. Preventing Chronic Disease **18** (2021). https://doi.org/10.5888/pcd18.200531, https://doi.org/10.5888/pcd18.200531
2. USDA: What we eat in America (WWEIA) database | Ag Data Commons 2023 (Mar 2023), https://data.nal.usda.gov/dataset/what-we-eat-america-wweia-

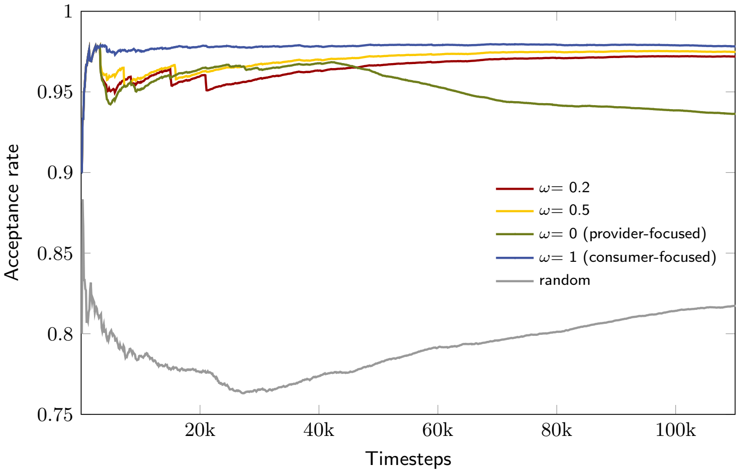


Fig. 6: Acceptance rates. Weighted models and the consumer-focused model achieve an increasing acceptance rate while the provider-focused model does not.

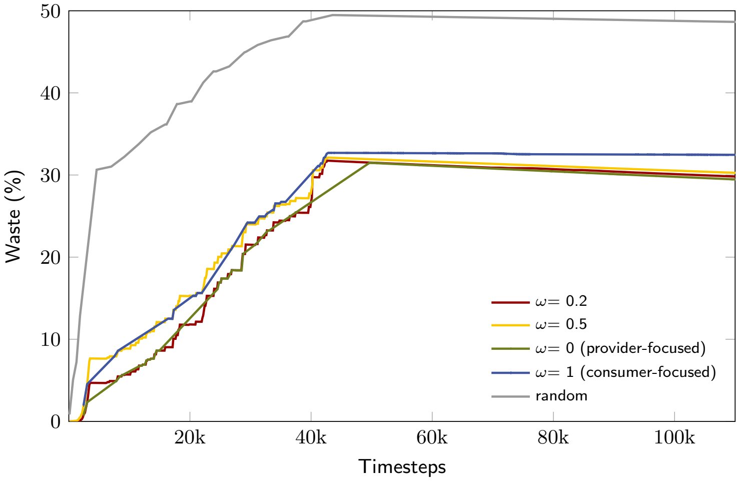


Fig. 7: Food waste tendency

database, [Accessed 2023-12-13]

2.2 AKIBoards: A Structure-Following Multiagent System for Predicting Acute Kidney Injury

AKIBoards: A Structure-Following Multiagent System for Predicting Acute Kidney Injury

DAVID GORDON[1][0000-0002-5273-8344], PANAYIOTIS PETOUSIS[1][0000-0002-0696-608X], SUSANNE B. NICHOLAS [1][0000-0003-3535-9120], ALEX A.T. BUI[1][0000-0002-4702-1373]

[1] University of California, Los Angeles
d.gordon@ucla.edu

Abstract. Diagnostic reasoning entails a physician's local (mental) model based on an assumed or known shared perspective (global model) to explain patient observations with evidence assigned towards a clinical assessment. But in several (complex) medical situations, multiple experts work together as a team to optimize health evaluation and decision-making by leveraging different perspectives. Such consensus-driven reasoning reflects individual knowledge contributing toward a broader perspective on the patient. In this light, we introduce *STRUCture-following for Multiagent Systems (STRUC-MAS),* a framework automating the learning of these global models and their incorporation as prior beliefs for agents in multiagent systems (MAS) to follow. We demonstrate proof of concept with a prosocial MAS application for predicting acute kidney injuries (AKIs). In this case, we found that incorporating a global structure enabled multiple agents to achieve better performance (average precision, AP) in predicting AKI 48 hours before onset (structure-following-fine-tuned, SF-FT, AP=0.195; SF-FT-retrieval-augmented generation, SF-FT-RAG, AP=0.194) vs. baseline (non-structure-following-FT, NSF-FT, AP=0.141; NSF-FT-RAG, AP=0.180) for balanced precision-weighted-recall-weighted voting. Markedly, SF-FT agents with higher recall scores reported lower confidence levels in the initial round on true positive and false negative cases. But after explicit interactions, their confidence in their decisions increased (suggesting reinforced belief). In contrast, the SF-FT agent with the lowest recall decreased its confidence in true positive and false negative cases (suggesting a new belief). This approach suggests that learning and leveraging global structures in MAS is necessary prior to achieving competitive classification and diagnostic reasoning performance.

Keywords: Multiagent Systems, Machine Learning, Structure Learning, Health Informatics.

1 Introduction

When performing rounds in a hospital, physicians evaluate patients and communicate their clinical reasoning with their peers, who may agree or disagree with their assessment. Similarly, when physicians participate in clinical boards (e.g., tumor board for reviewing potential cancers and treatment), experts collaboratively form an assessment and assign a diagnosis, with the goal of exploring the full range of possibilities. Both types of clinical interactions involve multiple perspectives that are used to determine a patient's "next step." Such approaches extend to other interdisciplinary scenarios, where generalists (e.g., internists) often consult with others for specific expertise (e.g., nephrologists) [1]. These approaches aim to optimize health outcomes and diagnostic reasoning by employing different types of experience, coalescing individual insights into a more complete picture of the patient that informs clinical reasoning [2]. By analogy, local knowledge is woven together into an explanation for a patient based on a shared knowledgebase (a "global structure") representing all patients. These collective assessments have been shown to improve patient care and outcomes in myriad settings [3-5].

However, despite its high value, there can be significant costs and difficulties associated with this approach [6]: not all healthcare environments have the same degree of access to expertise (e.g., a limited number or no experts in a low-resource setting) [1, 4]; communication may be challenging (e.g., due to lack of (quality) documentation to understand another physician's reasoning); and decision-making may be time sensitive [7]. Indeed, the use of clinical boards is underutilized in part because of the need for considerable human resources.

To help address these issues, we introduce the first implementation of *STRUCture-following for Multiagent Systems* (STRUC-MAS) [8], a framework that draws inspiration from the construct of a clinical group of experts in a prosocial context [9] to facilitate "on-demand" specialist knowledge workers. Specifically, STRUC-MAS provides a way to learn the "global structure" of a domain problem, which can then be incorporated as prior beliefs for agents in multiagent systems (MAS) to follow. To demonstrate, we address the critical problem of predicting the onset of acute kidney injury (AKI) as a group of interacting agents, each with different perspectives (i.e., AKIBoards) that update their beliefs over time to reach agreement in an explainable manner [8].

2 Methods

2.1 AKIBoards: Predicting acute kidney injuries

Fig. 1 overviews the general architecture of STRUC-MAS [8], which was adapted to the specific problem of predicting acute kidney injuries. From a health perspective, AKIs can lead to chronic kidney disease (CKD) and other long-term health complications. Most AKIs are preventable if there is sufficient time to intervene and provide treatment. However, the onset of many AKIs go undetected until too late, and some are not even recognized until after the fact. To address this problem, AKIBoards was designed as a data-driven, clinically-oriented board using the STRUC-MAS framework. Following Stein et al.'s definition of a citizen-centric multiagent system, we shaped AKIBoards to be: 1) citizen-aware, taking into account stakeholder (health system, clinician) perspectives and requirements; 2) citizen-beneficial, providing value by improving current processes positively; and citizen-auditable, provide interpretable/explainable output and allow input from stakeholders [10].

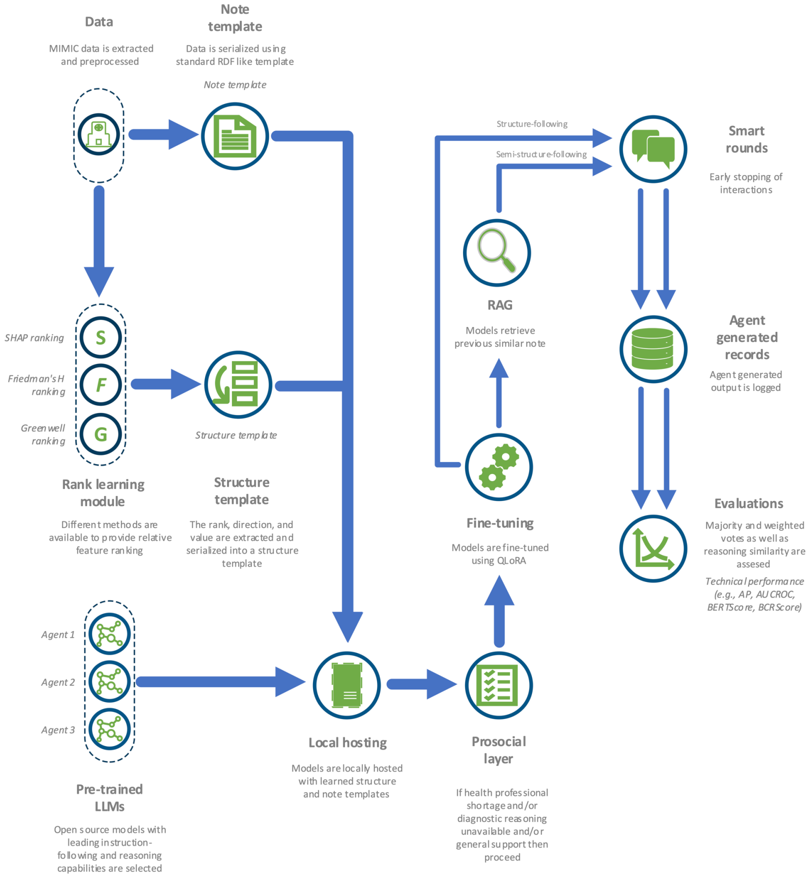

Data
MIMIC data is extracted and preprocessed
Note template
Data is serialized using standard RDF like template
Note template
Structure-following
Semi-structure-following
Smart rounds
Early stopping of interactions
SHAP ranking
S
Friedman's H ranking
F
Greenwell ranking
G
Structure template
RAG
Models retrieve previous similar note
Agent generated records
Agent generated output is logged
Rank learning module
Different methods are available to provide relative feature ranking
Structure template
The rank, direction, and value are extracted and serialized into a structure template
Fine-tuning
Models are fine-tuned using QLoRA
Evaluations
Majority and weighted votes as well as reasoning similarity are assesed
Technical performance (e.g., AP, AUCROC, BERTScore, BCRScore)
Agent 1
Agent 2
Agent 3
Local hosting
Models are locally hosted with learned structure and note templates
Prosocial layer
If health professional shortage and/or diagnostic reasoning unavailable and/or general support then proceed
Pre-trained LLMs
Open source models with leading instruction-following and reasoning capabilities are selected

Fig. 1. AKIBoards in STRUC-MAS. Models are locally hosted with note and structure template. If the criteria from the prosocial layer are met, the framework proceeds with structure-following or semi-structure-following. Agent generated output is logged for evaluation.

Dataset and representation. To develop AKIBoards, an AKI dataset was extracted from MIMIC-III [11], a freely-available database comprising deidentified health-related data [11]. Details regarding the dataset as well as the developed AKI algorithm can be found in a previous study [12]. Briefly, the dataset contains clinical laboratory values and was stratified and randomized into train (70%, n=9,176), valid (15%, n=1,966), and test (15%, n=1,967) sets. The fine-tuning of a pre-trained large language model (LLM) was fit on the training set, the structures fit on the validation set, and the agents evaluated on the holdout test set. We adapted a basic note template, where the data is serialized into knowledge triples (*feature name*, *is*, *value*) [13]. Notably, serialization transforms the tabular data into a standardized note.

Structure learning: Rank learning module and structure template. When the structure is unknown to clinicians and/or agents, traditional machine learning or statistics methods can help learn plausible structures to follow [12]. We build on the rank learning module from Ranking Approaches for Unknown Structures (RAUS) [14] to learn structures that also capture feature value directionality and pairwise interactions [15-17]. In this work, we used SHapley Additive exPlanations (SHAP) given its wide acceptance as an interpretability method. We transformed categorical features into dummy variables to enable interpretability at the bin level. Further, we used best-k (i.e., top) ranked features, where k=10 (Appendix 6.1, Figure 2). Note that SHAP output is automatically incorporated into the standardized structure template (see Appendix 6.1, Table 1).

Prosocial layer. We integrate prosocial logic [9] via Boolean constraints to help ensure that the system is used for augmenting physicians rather than replacing them. The key issues this system aims to address are: 1) health professional shortages, either due to time constraints or unavailability; 2) unavailable reasoning for an AKI diagnosis, or to further support clinical decision making; and/or 3) general support (e.g., training of physicians to detect AKIs, demonstrating the reasoning process). These issues are incorporated into a "prosocial" score (PScore, Eq. 1) to measure the complementary nature of the implementation. Note that the system must address at least one issue to have permission to run.

$$PScore\ =\ i_1\alpha_1 + \cdots + i_{\mathrm{n}}\alpha_n\ \ (1)$$

where i_1 is issue 1 (health professional shortages), α_1 is the weight of i_1, i_2 is a secondary issue (unavailable reasoning), α_2 is the weight of i_2, etc. Two options are available for each issue (true or false), where the value (ranges from 0-1) for true is 0.99 and false is 0.01. Note that the higher the value the more "prosocial" the option. Further, α_1, …, α_n, must sum to 1. In this work, all issues are equally weighted (i.e., 0.333). Thus, the minimum PScore required for the system to run is 0.336. As we identified and addressed three issues (options set to true), the PScore was 0.989. This approach can be used to address a single or multiple issues.

Pre-trained LLMs, fine-tuning, and retrieval-augmented generation. LLMs in STRUC-MAS exploit the global structure (Appendix 6.1, Table 1) to infer local structures (i.e., individualized diagnostic reasoning) and diagnoses over time (i.e., iterations of iteration). We designed two paths: 1) *structure-following*, which exploits the global structure; and 2) *semi-structure-following*, which exploits the global structure and explores via retrieval-augmented generation (RAG) [18] – enabling agents to retrieve a previous similar note from the training set for comparison. Three open-source LLMs ranging from 8-32B parameters were selected based on their reported leading instruction-following and reasoning capabilities: 1) QWEN 2.5 Instruct [19]; Phi4 [20]; and Llama 3.1 Instruct [21], and fine-tuned via quantized low-rank adaptation (QLoRA) [22].

Smart rounds. To efficiently orchestrate multiple LLMs we build on existing frameworks and design a coordination style that is more akin to health system routines (see Appendix 6.2, Fig. 4) [23]. This setup mimics clinical rounds, where there is frequently a mixture of expertise levels (e.g., medical students, residents, attendings), to teach less experienced experts to perform at the level of more experienced experts (i.e., a mixture of experts) [24]. In this scenario, the assumption is that individual weaker experts can adopt the knowledge from stronger/more experienced experts to update their beliefs over time/rounds (i.e., knowledge distillation) [25]. Further, stronger experts can reinforce and/or increase their confidence in their beliefs. Likewise, smart rounds aim to optimize reaching consensus across the agents in as few rounds as possible without sacrificing performance. If performance gain in a given round is less than a set threshold then early stopping occurs (i.e., the agents achieved near maximum exploitation of the global structure).

Multi-agent system logs/records and evaluation. Multiagent records (MAR) log agent-generated output using a new vocabulary called agent-based terms (ABT) [26]. (see Appendix 6.3, Table 2-3). We evaluated the agent-based diagnosis (AD) using confusion matrix-based metrics (area under receiver operating characteristic curve, AUCROC; average precision, AP; precision; recall; false positive, FP; false negative, FN; true positive, TP; true negative, TN). We evaluated the agent-based diagnostic reasoning (ADR) using semantic similarity metrics (BERTScore) [27]. We developed a balanced classification and reasoning score (BCRScore, Eq. 2) to jointly assess AD and ADR (see Appendix 6.8, Table 8):

$$\text{BCRScore} = \text{A}\alpha + \text{B}\beta \quad (2)$$

where A is the AD metric (e.g., AP), α is A's weight, B is the ADR metric (e.g., average BERTScore F1), and β is B's weight. Note that α and β must sum to 1. Further, we assess agent confidence levels (ACL) as well as agent documentation burden via agent time spent on documentation (ATSD) and agent documentation length (ADL).

3 Results

Round 0 shows SF-FT Agents 2 and 3 perform similarly with AP of 0.186 and 0.202, respectively (see Appendix 6.4, Table 4). Notably, in Round 0, SF-FT Agent 1 performed poorly (AP of 0.133) with a recall of 0.01 (misdiagnosing almost all TP as FN), whereas SF-FT Agents 2 and 3 achieved a recall of 0.67 and 0.62, respectively. Round 1 shows SF-FT Agents 1, 2, and 3 perform similarly (AP of 0.190, 0.192, 0.198, respectively). Interestingly, via explicit interactions, SF-FT Agents 1 and 2 recall increased to 0.62 and 0.68, respectively (i.e., knowledge distillation). Similarly, SF-FT-RAG Agent 1 and Agent 2 recall increased to 0.52 and 0.69, respectively. Further, the recall for the BPRV increased from 0.37 to 0.62 in Round 1 for SF-FT (the team captures approximately 1.6 TP for every FN). Note that since P (round 1 BPRV AP) – O (round 0 BPRV AP) < Q (0.040), early stopping occurred.

Markedly, SF-FT Agent 1 was highly confident in Round 0 regarding TP and FN cases even though it was misdiagnosing almost all positive cases as negative cases, but after explicit interactions, its high confidence in TP and FN cases dropped (see Appendix 6.6, Table 6). We also observed that Agents 2 and 3 increased their confidence in TP and FN cases from Round 0 to Round 1, suggesting the explicit interactions reinforced their prior beliefs (see Appendix 6.6, Table 6). Appendix Table 7 shows that incorporating RAG made all agents highly confident in Round 0 TP and FN cases. Appendix Figure 5-6 shows the ADR alignment analysis by TP, FP, FN, and TN cases, highlighting that SF-FT Agent 3 reference group SF-FT Agent 2 were more aligned than SF-FT Agent 1 reference group SF-FT 2 in Round 0, while in Round 1 the alignment increased. Appendix Figure 7-8 shows that the SF-FT-RAG Agent 2 ADR was more aligned with SF-FT Agent 2 for TP, FP, FN, and TN cases. Appendix Table 5 shows the agent documentation burden. Appendix Table 8 shows the BCRScores.

4 Discussion

In this work, we highlight the role global structure plays in LLM-based agents. In our use case for predicting AKI, pretraining, fine-tuning, and/or RAG are no replacement for learning structure via traditional machine learning or statistics methods. In reviewing the “reasoning” for structure-following agents, we found that it leveraged the global structure of the data and associations between variables to provide local answers. Board configurations may be suitable for other clinical specialties as well, such as where the overlap of organs/systems makes it difficult to assume or know the underlying structures (e.g., oncology, cardiology, endocrinology, gastroenterology, rheumatology). Structure-following approaches seem like a promising path forward to utilize and evaluate these emerging technologies in the health domain. Future work may explore implementing multiple boards (i.e., multi-stakeholders) and dynamic resource allocation.

5 Conclusion

We demonstrated that global structure is necessary prior to achieving competitive classification and reasoning performance in the health domain. Further, we showed that not all models can leverage the global structure in a meaningful way; however, models that are stronger in specific capabilities can help the team improve. We showed that across multiple rounds agents can increase and decrease their confidence levels based on explicit interactions with other models. We also demonstrated that smart rounds were sufficient to get cheaper, faster, and informative results.

Author Contributions. Most intellectual work was done by DG with guidance from AB. DG planned and implemented the paper methodologies. DG created the software and multiagent records as well as ran the experiments and evaluated the models. PP, SN, and AB reviewed drafts of the paper and provided feedback.

References

[1] N. Bansal, N. Arora, D. Mariuma, J. A. Jefferson, K. O'Brien, and S. Shankland, "Mission and 1-year outcomes of a cardiorenal subspecialty consultation service," *Kidney360,* vol. 3, no. 4, pp. 749-751, 2022.

[2] J. P. Kassirer, "Diagnostic reasoning," *Annals of internal medicine,* vol. 110, no. 11, pp. 893-900, 1989.

[3] E. A. Newman *et al.*, "Changes in surgical management resulting from case review at a breast cancer multidisciplinary tumor board," *Cancer,* vol. 107, no. 10, pp. 2346-2351, 2006.

[4] N. S. El Saghir, N. L. Keating, R. W. Carlson, K. E. Khoury, and L. Fallowfield, "Tumor boards: optimizing the structure and improving efficiency of multidisciplinary management of patients with cancer worldwide," *American Society of Clinical Oncology Educational Book,* vol. 34, no. 1, pp. e461-e466, 2014.

[5] N. S. El Saghir *et al.*, "Survey of utilization of multidisciplinary management tumor boards in Arab countries," *The Breast,* vol. 20, pp. S70-S74, 2011.

[6] W. L. Kissick, *Medicine's dilemmas: infinite needs versus finite resources*. Yale University Press, 1994.

[7] M. H. Murad *et al.*, "Measuring Documentation Burden in Healthcare," *Journal of general internal medicine,* pp. 1-12, 2024.

[8] D. Gordon, "STRUC-MAS: STRUCture-following for MultiAgent Systems," ed: Zenodo, 2025.

[9] F. P. Santos, "Prosocial dynamics in multiagent systems," *AI Magazine,* vol. 45, no. 1, pp. 131-138, 2024.

[10] S. Stein and V. Yazdanpanah, "Citizen-centric multiagent systems," 2023.

[11] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," (in eng), *Sci Data,* vol. 3, p. 160035, May 2016, doi: 10.1038/sdata.2016.35.

[12] D. Gordon *et al.*, "Automated Dynamic Bayesian Networks for Predicting Acute Kidney Injury Before Onset," *arXiv preprint arXiv:2304.10175,* 2023, doi: 10.48550/arXiv.2304.10175.

[13] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, "Tabllm: Few-shot classification of tabular data with large language models," in *International Conference on Artificial Intelligence and Statistics*, 2023: PMLR, pp. 5549-5581.

[14] D. Gordon, "RAUS: Ranking Approaches for Unknown Structure Learning," ed: Zenodo, 2023.

[15] S. Lundberg, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874,* 2017.

[16] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," 2008.

[17] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy, "A simple and effective model-based variable importance measure," *arXiv preprint arXiv:1805.04755,* 2018.

[18] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems,* vol. 33, pp. 9459-9474, 2020.

[19] A. Yang *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115,* 2024.

[20] M. Abdin *et al.*, "Phi-4 technical report," *arXiv preprint arXiv:2412.08905,* 2024.

[21] A. Dubey *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783,* 2024.

[22] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in Neural Information Processing Systems,* vol. 36, 2024.

[23] Q. Wu *et al.*, "Autogen: Enabling next-gen llm applications via multi-agent conversation framework," *arXiv preprint arXiv:2308.08155,* 2023.

[24] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation,* vol. 3, no. 1, pp. 79-87, 1991.

[25] G. Hinton, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531,* 2015.

[26] D. Gordon, "Multiagent Records [Database]," ed. Zenodo, 2025.

[27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675,* 2019.

6 Appendix

6.1 Learning the Structure Template

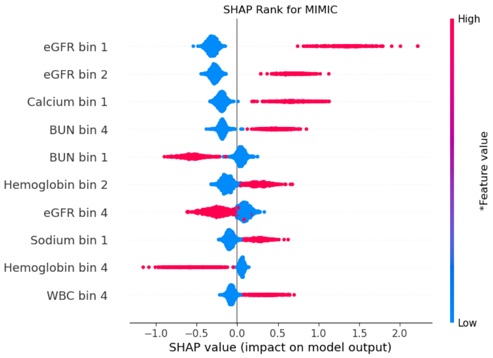


Fig. 2. SHAP value feature rank for MIMIC.

Table 1. Autogenerated Structure Template

Site	Structure Template
MIMIC	Having the lowest bin (i.e., 1) for estimated glomerular filtration rate (eGFR) is the most important feature and indicates the highest risk for acute kidney injury (AKI). Having the second lowest bin (i.e., 2) for eGFR is the second most important feature and indicates higher risk for AKI. Having the lowest bin for calcium (i.e., 1) is the third most important feature and indicates higher risk for AKI. Having the highest bin for blood urea nitrogen (i.e., 4) is the fourth most important feature and indicates higher risk for AKI. Having the lowest bin for blood urea nitrogen (i.e., 1) is the fifth most important feature and indicates decreased risk for AKI. Having the second lowest bin for hemoglobin (i.e., 2) is the sixth most important feature and indicates higher risk for AKI. Having the highest bin for eGFR (i.e., 4) is the seventh most important feature and indicates decreased risk for AKI. Having the lowest bin for sodium (i.e., 1) is the eighth most important feature and indicates higher risk for AKI. Having the highest bin for hemoglobin (i.e., 4) is the ninth most important feature and indicates decreased risk for AKI. Having the highest bin for white blood cell count (i.e., 4) is the tenth most important feature and indicates increased risk for AKI.

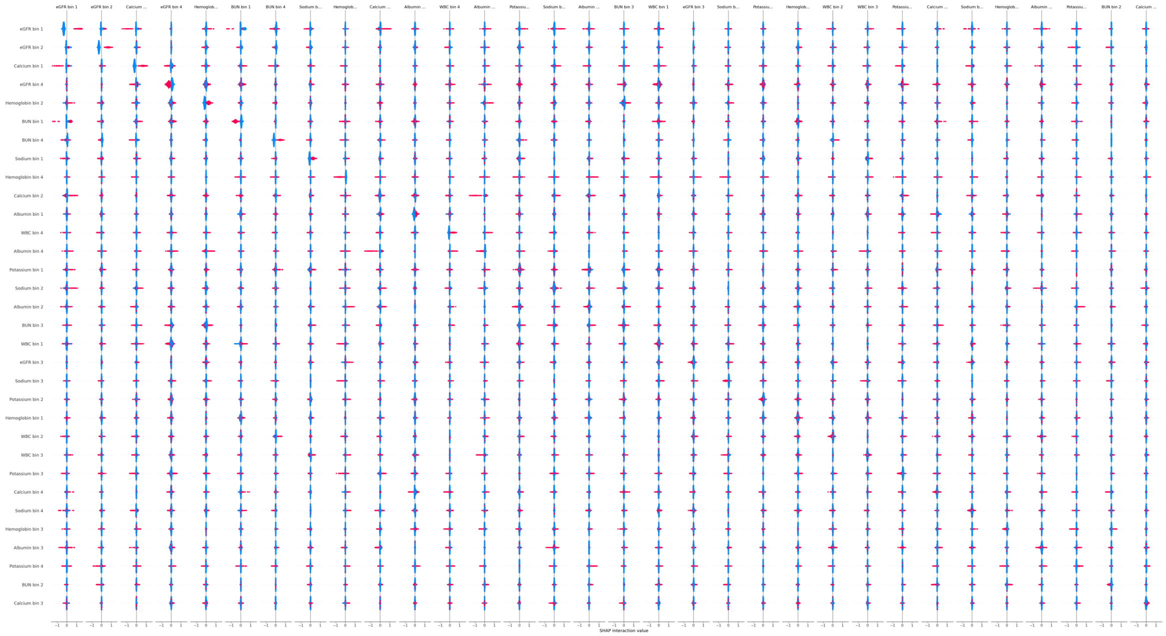


Fig. 3. SHAP interaction values for MIMIC. Main effects are on the diagonal and interaction effects are off-diagonal.

6.2 Smart Rounds

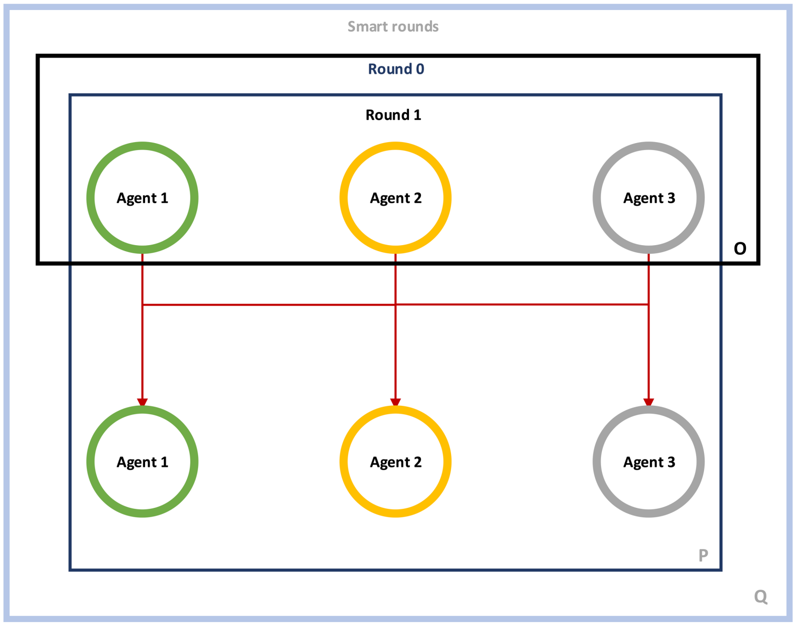


Fig. 4. Smart rounds: Include an outer plate (Round 0) and an inner plate (Round 1). Note that the inner plate repeats until P-O < Q. Q is the smart rounds early stopping threshold (e.g., Q = 0.040), P is the round 1 AP, and O is round 0 AP. In round 0 (implicit interactions), experts (i.e., agents) do their own individual assessments, and the end user (e.g., healthcare provider) may then calculate the aggregate/population statistics across the multiple agents via majority vote, precision-weighted vote, recall-weighted vote, or balanced precision-weighted-recall-weighted vote (BPRV). In round 1+ (explicit interactions), agents explicitly interact/consult with each other via conversational interactions (directed edges) to potentially update their prior diagnosis and prior diagnostic reasoning from round 0. This results in updated diagnosis, diagnostic reasoning and aggregate statistics.

6.3 Multi-Agent Records Agent-Based Terms (MAR-ABT)

Table 2. Multiagent records agent-based terms. Consists of standard terms, such as Agent Name (AN), Agent Diagnosis (AD), Agent Diagnostic Reasoning (ADR), Agent Confidence Level (ACL), Agent Documentation Length (ADL), and Agent Time Spent on Documentation (ATSD).

	Agent Name (AN)	**Agent Diagnosis (AD)**	**Agent Diagnostic Reasoning (ADR)**	**Agent Confidence Level (ACL)**	**Agent Documentation Length (ADL)**	**Agent Time Spent on Documentation (ATSD)**
Reference	Structure-Following Agent 1	1.01, AKI	The patient has several indicators that suggest a higher risk for acute kidney injury (aki). the egfr is in the second lowest bin (2), indicating a higher risk. additionally, hemoglobin is in the second lowest bin (2) and sodium is in the lowest bin (1), both of which also indicate higher risk. while some factors like white blood cell count being in the third bin do not strongly suggest aki, the combination of multiple high-risk indicators suggests that there's a moderate likelihood of developing aki within two days.	High	1521 tokens	11 Seconds
Reference	Structure-Following Agent 2	1.01, AKI	The patient has multiple indicators (egfr at 2, hemoglobin at 2, sodium at 1, and calcium at 1) that are associated with a higher risk for acute kidney injury. these factors collectively suggest an increased likelihood of developing aki within two days.	High	1872 tokens	43 Seconds
Reference	Structure-Following Agent 3	1.01, AKI	The patient has several key indicators that suggest a higher risk for acute kidney injury (aki). the hemoglobin is in the second lowest bin (2), indicating a higher risk. additionally, sodium and calcium are both in their respective lowest bins (1), which also indicate higher risk. these factors collectively increase the likelihood of developing aki within two days.	High	1486 tokens	44 Seconds

Table 3. Mapping (source) agent-based terms (ABT) vocabulary to (target) other standard vocabularies

Vocabulary Mapping	**Source**	**Target**
AN → PN	Structure_Following_Agent_2	Smith, Jane MD
AD → ICD-9	AD: 1.01	ICD-9: 589.4
AD → ICD-10	AD: 1.01	ICD-10: N17.9
AD → SNOMED	AD: 1.01	SNOMED: 140031000119103
ADR → Clinical Note (CN)	The patient has multiple indicators (egfr at 2, hemoglobin at 2, sodium at 1, and calcium at 1) that are associated with a higher risk for acute kidney injury. these factors collectively suggest an increased likelihood of developing aki within two days.	The patient's chief complaint was weakness. labs show low egfr, sodium, and calcium. Monitor labs for changes. Potential electrolyte imbalance, consult nephrology.

6.4 Agent Diagnosis Performance Evaluations

Table 4. Structure-following agents vs. Non-structure-following agents (baselines)

Round	Models	Structure-following fine-tuned (SF-FT)				Structure-following fine-tuned retrieval-augmented generation (SF-FT-RAG)				Non-structure following fine-tuned (NSF-FT baseline)				Non-structure following fine-tuned retrieval-augmented generation (NSF-FT-RAG baseline)			
		AUC[h]	AP	Pr.[d]	Re.[e]	AUC[h]	AP	Pr.[d]	Re.[e]	AUC[h]	AP	Pr.[d]	Re.[e]	AUC[h]	AP	Pr.[d]	Re.[e]
0	Agent 1[a]	0.504	0.133	0.30	**0.01**	0.552	0.153	0.242	**0.20**	0.50	0.131	nan	**0.00**	0.500	0.131	nan	**0.0**
	Agent 2[b]	0.647	0.186	0.21	**0.67**	0.673	0.201	0.232	**0.69**	0.55	0.144	0.150	**0.72**	0.618	0.173	0.205	**0.57**
	Agent 3[c]	0.666	0.202	0.25	**0.62**	0.667	0.213	0.282	**0.54**	0.50	0.131	nan	**0.00**	0.619	0.170	0.230	**0.48**
	Majority Vote	0.659	0.202	0.26	**0.57**	0.662	0.208	0.272	**0.54**	0.50	0.131	nan	**0.00**	0.632	0.200	0.30	**0.41**
	Precision-Weighted Vote[f]	0.505	0.135	0.43	**0.01**	0.556	0.167	0.371	**0.15**	0.50	0.131	nan	**0.00**	0.50	0.1312	nan	**0.00**
	Recall-Weighted Vote[g]	0.654	0.187	0.21	**0.72**	0.673	0.201	0.232	**0.69**	0.55	0.144	0.150	**0.72**	0.605	0.165	0.183	**0.64**
	Balanced Precision-Weighted-Recall-Weighted Vote[i]	**0.580**	**0.161**	**0.32**	**0.37**	**0.615**	**0.184**	**0.302**	**0.42**	**0.53**	**0.138**	**nan**	**0.36**	**0.553**	**0.148**	**nan**	**0.32**
1	Agent 1 [a]	0.649	0.190	0.23	**0.62**	0.625	0.180	0.225	**0.52**	0.53	0.139	0.144	**0.60**	0.586	0.169	0.255	**0.31**
	Agent 2 [b]	0.659	0.192	0.22	**0.68**	0.672	0.201	0.232	**0.69**	0.54	0.143	0.152	**0.57**	0.631	0.181	0.217	**0.58**
	Agent 3 [c]	0.652	0.198	0.25	**0.55**	0.665	0.208	0.269	**0.56**	0.50	0.131	0.000	**0.00**	0.633	0.201	0.303	**0.41**
	Majority Vote	0.648	0.190	0.23	**0.61**	0.662	0.200	0.246	**0.60**	0.52	0.137	0.143	**0.49**	0.637	0.195	0.268	**0.47**
	Precision-Weighted Vote[f]	0.652	0.198	0.25	**0.55**	0.628	0.189	0.260	**0.45**	0.52	0.137	0.143	**0.49**	0.583	0.178	0.328	**0.24**
	Recall-Weighted Vote[g]	0.661	0.192	0.22	**0.69**	0.671	0.198	0.225	**0.71**	0.55	0.145	0.151	**0.69**	0.632	0.181	0.217	**0.58**
	Balanced Precision-Weighted-Recall-Weighted Vote[i]	**0.656**	**0.195**	**0.24**	**0.62**	**0.650**	**0.194**	**0.243**	**0.58**	**0.54**	**0.141**	**0.147**	**0.59**	**0.608**	**0.180**	**0.272**	**0.41**

[a] Meta Llama 3.1 8B Parameters; [b] Microsoft Phi4 14B Parameters; [c] Alibaba Cloud Qwen 2.5 IT 32B Parameters;

[d] Precision = TP/(TP+FP); [e] Recall = TP/(TP+FN); [f] Threshold set at 0.75; [g] Threshold set at 0.25; [h] AUCROC; [i] (Precision-weighted vote + Recall-weighted vote) /2.

6.5 Agent Documentation Burden

Table 5. SF-FT Agent Time Spent on Documentation (ATSD) and Agent Documentation Length (ADL)

Round	**Models**	**Agent Time Spent on Documentation (MM:SS.mmm or Min:Sec.Msec)**					**Agent Documentation Length (Total tokens)**				
		Min.	**25%**	**50%**	**75%**	**Max.**	**Min.**	**25%**	**50%**	**75%**	**Max.**
0	Agent 1	0.0017	8.476	**10.640**	11.961	18.188	570	642	**716**	758	975
	Agent 2	0.0025	28.593	**30.680**	32.481	49.646	783	907	**937**	966	1196
	Agent 3	0.0004	34.302	**39.579**	1:13.494	2:40.391	578	617	**630**	645	731
1	Agent 1	0.0024	9.864	**11.327**	13.531	23.612	1173	1360	**1461**	1567	2018
	Agent 2	0.0018	37.221	**40.428**	43.616	58.986	1406	1697	**1792**	1883	2235
	Agent 3	0.0010	46.841	**53.270**	1:23.257	3:07.440	1186	1366	**1440**	1521	1830

In Table 5 above we see the median agent time spent on documentation ranges from 11-40 seconds in round 0 to 11-53 seconds in round 1. Further, we see the median agent documentation length ranged from 630-937 tokens in round 0 to 1440-1792 tokens in round 1.

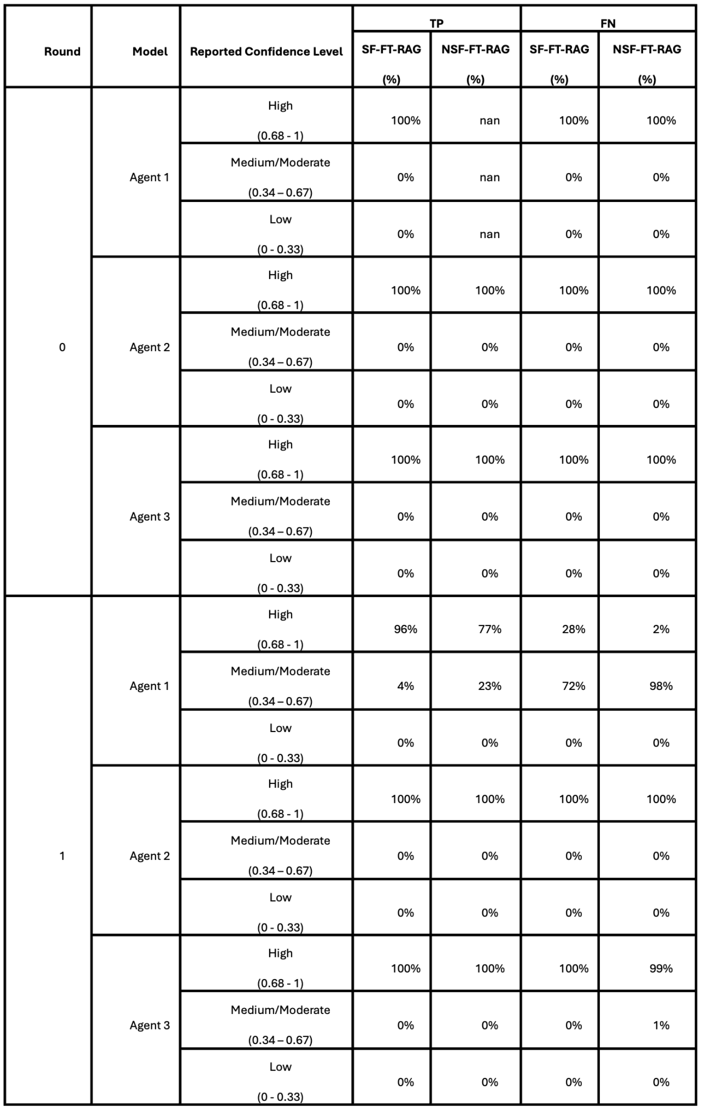
6.6 Agent Confidence Levels

Table 6. SF-FT Agents reported confidence levels by TP, FP, TN, and FN.

Round	Model	Reported Confidence Level	TP (%)	FP (%)	TN (%)	FN (%)
0	Agent 1	High (0.68 - 1)	100%	57%	91%	96%
		Medium/Moderate (0.34 – 0.67)	0%	43%	9%	4%
		Low (0 - 0.33)	0%	0%	0%	0%
	Agent 2	High (0.68 - 1)	64%	42%	59%	46%
		Medium/Moderate (0.34 – 0.67)	36%	58%	41%	53%
		Low (0 - 0.33)	0%	0%	0%	1%
	Agent 3	High (0.68 - 1)	51%	23%	0%	0%
		Medium/Moderate (0.34 – 0.67)	49%	77%	10%	11%
		Low (0 - 0.33)	0%	0%	90%	89%
1	Agent 1	High (0.68 - 1)	86%	80%	87%	88%
		Medium/Moderate (0.34 – 0.67)	14%	20%	10%	11%
		Low (0 - 0.33)	0%	0%	3%	1%
	Agent 2	High (0.68 - 1)	97%	91%	94%	89%
		Medium/Moderate (0.34 – 0.67)	3%	9%	6%	11%
		Low (0 - 0.33)	0%	0%	0%	0%
	Agent 3	High (0.68 - 1)	94%	91%	63%	58%
		Medium/Moderate (0.34 – 0.67)	6%	9%	37%	42%
		Low (0 - 0.33)	0%	0%	0%	0%

Table 6 above shows that in the initial round Agent 1 was highly confident in its diagnosis and ADR of FN cases; however, Agent 1 missed almost all the positive cases, suggesting it couldn't differentiate between the groups very effectively. After explicit interactions in round 1 we see that Agent 1 reduced its high confidence, suggesting the other agent's diagnosis and ADR was influential. Further, we see that in the initial round Agent 2 and Agent 3, though having considerably higher recall than Agent 1, were more uncertain and after explicit interactions with the other agent's their confidence levels increased.

Table 7. SF-FT-RAG and NSF-FT-RAG Agents reported confidence levels by TP and FN.



Round	Model	Reported Confidence Level	TP		FN	
			SF-FT-RAG (%)	NSF-FT-RAG (%)	SF-FT-RAG (%)	NSF-FT-RAG (%)
0	Agent 1	High (0.68 - 1)	100%	nan	100%	100%
		Medium/Moderate (0.34 – 0.67)	0%	nan	0%	0%
		Low (0 - 0.33)	0%	nan	0%	0%
	Agent 2	High (0.68 - 1)	100%	100%	100%	100%
		Medium/Moderate (0.34 – 0.67)	0%	0%	0%	0%
		Low (0 - 0.33)	0%	0%	0%	0%
	Agent 3	High (0.68 - 1)	100%	100%	100%	100%
		Medium/Moderate (0.34 – 0.67)	0%	0%	0%	0%
		Low (0 - 0.33)	0%	0%	0%	0%
1	Agent 1	High (0.68 - 1)	96%	77%	28%	2%
		Medium/Moderate (0.34 – 0.67)	4%	23%	72%	98%
		Low (0 - 0.33)	0%	0%	0%	0%
	Agent 2	High (0.68 - 1)	100%	100%	100%	100%
		Medium/Moderate (0.34 – 0.67)	0%	0%	0%	0%
		Low (0 - 0.33)	0%	0%	0%	0%
	Agent 3	High (0.68 - 1)	100%	100%	100%	99%
		Medium/Moderate (0.34 – 0.67)	0%	0%	0%	1%
		Low (0 - 0.33)	0%	0%	0%	0%

6.7 Agent Diagnostic Reasoning Alignment Analysis

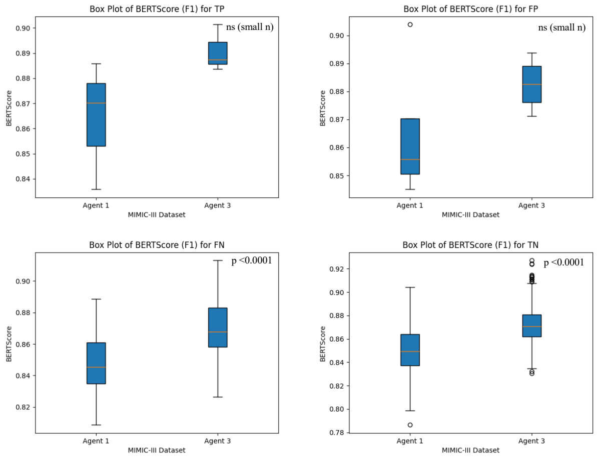


Fig. 5. SF-FT Round 0

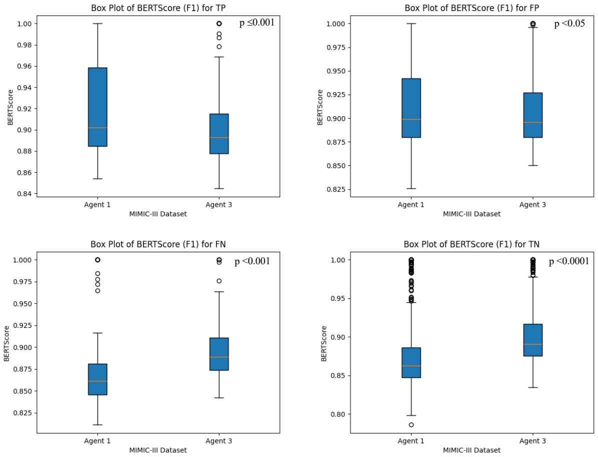


Fig. 6. SF-FT Round 1

Fig. 5-6. In Fig. 5 above we see Agent 3 reference group Agent 2 (highest recall) is more aligned than Agent 1 reference group Agent 2. Since Agent 2 and Agent 3 have similar precision and recall scores we expect those two agents to be more aligned across case types. In Fig. 6 we see that after explicit interactions the BERTScore increases. Further, we see that the gap between Agent 3 reference group Agent 2 and Agent 1 reference group Agent 2 decreases for TP and FP cases, suggesting more similar ADR. Yet, they are still statistically significant different, suggesting their ADR are not identical (i.e., there exists some unique reasoning).

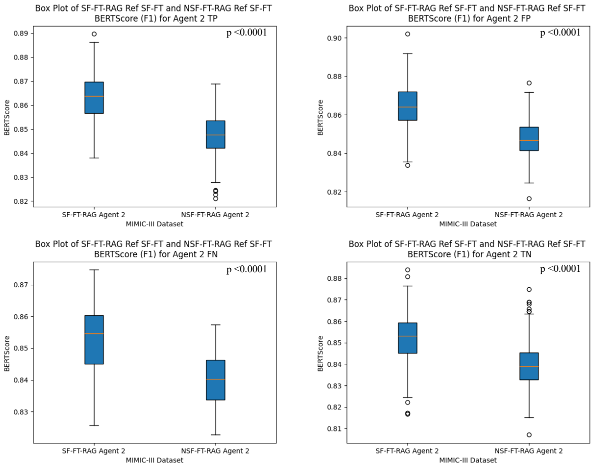


Fig. 7. SF-FT-RAG Agent 2 reference group SF-FT Agent 2 and NSF-FT-RAG Agent 2 reference group SF-FT Agent 2 Round 0

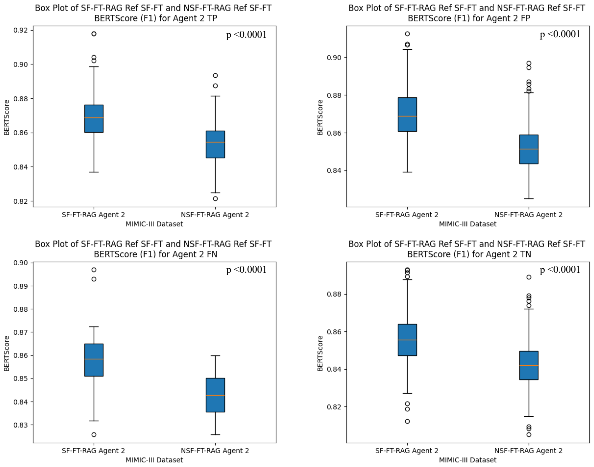


Fig. 8. SF-FT-RAG Agent 2 reference group SF-FT Agent 2 and NSF-FT-RAG Agent 2 reference group SF-FT Agent 2 Round 1

Fig. 7-8. In Fig. 7 above we see SF-FT-RAG Agent 2 reference group SF-FT Agent 2 (highest SF-FT recall agent) is more aligned than NSF-FT-RAG Agent 2 reference group SF-FT Agent 2. We hope to see this trend continue in subsequent rounds to demonstrate the utility of structure-following combined with RAG (i.e., semi-structure-following) vs. RAG alone. In Fig. 8 we see that after explicit interactions the BERTScores increase and SF-FT-RAG Agent 2 reference group SF-FT Agent 2 remains more aligned than NSF-FT-RAG Agent 2 reference group SF-FT Agent 2, demonstrating the benefit of leveraging the global structure.

6.8 Joint Assessment of Classification and Reasoning

Table 8. Balanced Classification and Reasoning Score (BCRScore)

AD Metric	α	ADR Metric	β	Model	Case Type	BCRScore	
						Round 0	Round 1
AP	0.5	Average BERTScore F1	0.5	SF-FT Agent 1 Ref. SF-FT Agent 2	FN	(0.161)*0.5+ (0.846)*0.5 =0.5035	(0.195)*0.5+ (0.878)*0.5 =0.5365
AP	0.5	Average BERTScore F1	0.5	SF-FT Agent 1 Ref. SF-FT Agent 2	TP	(0.161)*0.5+ (0.864)*0.5 =0.5125	(0.195)*0.5+ (0.919)*0.5 =0.5570
AP	0.5	Average BERTScore F1	0.5	SF-FT-RAG Agent 2 Ref. SF-FT Agent 2	FN	(0.184)*0.5+ (0.853)*0.5 = 0.5185	(0.194)*0.5+ (0.858)*0.5 =0.5260
AP	0.5	Average BERTScore F1	0.5	NSF-FT-RAG Agent 2 Ref. SF-FT Agent 2	FN	(0.148)*0.5+ (0.841)*0.5 =0.4945	(0.180)*0.5+ (0.843)*0.5 =.5115

The BCRScore aims to provide a balance between the agent classification performance and the agent diagnostic reasoning. In this implementation, we set α and β to 0.50. However, end-users may tune α and β to favor classification over reasoning and vice versa, as long as they sum to 1. Note that while we use the average precision (AP) of the balanced precision-weighted-recall-weighted vote for the AD metric, it can be replaced with other metrics (e.g., AUCROC, etc.). Also, regarding the ADR metric, we selected the average BERTScore F1, but it can be replaced with other metrics (e.g., (average) Rouge, (average) BLEU, etc.). Note that the higher the BCRScore the better the overall performance for diagnosis as well as diagnostic reasoning alignment. In Table 8 above we see that the BCRScore in round 1 was higher than in round 0 for SF-FT Agent 1 reference group SF-FT Agent 2 for FN cases, suggesting improvement in agent diagnosis and agent diagnostic reasoning for FN cases. Similarly, we see that the BCRScore in round 1 was higher than in round 0 for SF-FT Agent 1 reference group SF-FT Agent 2 for TP cases.

3 Modelling Human Needs in Shared Environments

3.1 EnvCraft: A Modular Environment Design Framework for Investigating Prosocial Incentives

EnvCraft: A Modular Environment Design Framework for Investigating Prosocial Incentives

Daniel E. Collins[0000−0002−1075−4063], Conor Houghton[0000−0001−5017−9473], and Nirav Ajmeri[0000−0003−3627−097X]

University of Bristol, Bristol, UK
{daniel.collins,conor.houghton,nirav.ajmeri}@bristol.ac.uk

Abstract. Developing approaches for ensuring that the collective behaviour of Multi-Agent Systems (MAS) is beneficial for all human stakeholders is a critical challenge for the safe deployment of MAS in real-world settings, where the goals of agent owners may partially conflict. A key direction towards addressing this challenge is understanding how to reliably incentivise prosocial behaviour among self-interested learning agents with partially conflicting goals through simulation experiments. However, methods for incentivising certain behaviours can lead to unintended consequences if the underlying influence of environmental and social pressures on behaviour are poorly understood. Avoiding side effects of behavioural incentives is challenging in the real world, where environmental and social pressures are dynamic and constantly changing.

In this paper, we present ongoing work developing EnvCraft, a framework for composing diverse multi-agent environments from modular building blocks that are parametrised to enable precise control and systematic variation of environment and population conditions. We detail requirements of EnvCraft for supporting robust evaluation of methods for incentivising prosocial behaviour, and for characterising the relationships between incentives, environment conditions and learned behaviour, through systematic exploration of the environment configuration parameters.

1 Introduction

Understanding how to design autonomous agents that can reliably achieve socially beneficial outcomes alongside other agents in multi-agent systems (MAS) is an important challenge towards the development of agents that can safely interact with humans and other artificial agents in the real world. Real-world multi-agent settings are inherently mixed-motive, i.e. there is partial alignment and conflict between the individual goals of agents arising from their assigned roles and the interests of human owners and stakeholders. Recent works investigating how different kinds of prosocial behaviour can be incentivised among self-interested learning agents present a promising direction for future research. Through simulation experiments in mixed-motive multi-agent environments, researchers have studied models of various human social mechanisms, such as indirect reciprocity [8], social norms and ethical principles [2, 10].

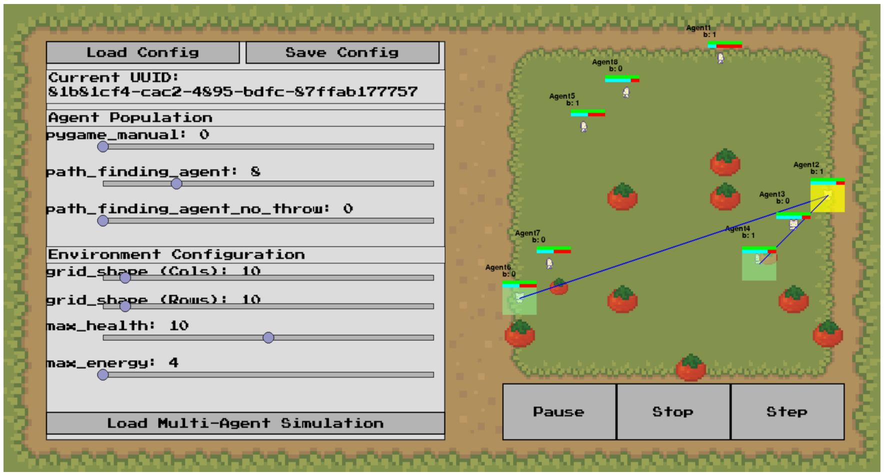
Specialised benchmarks and simulation frameworks are important across MAS research for supporting the development and evaluation of novel methods in key problem areas. While there are several popular frameworks for multi-agent reinforcement learning (MARL) that provide diverse mixed-motive multi-agent scenarios [5, 1, 9, 6], there is a need for specialised environments that better capture the challenges associated with incentivising prosocial behaviour in the real world. Real world multi-agent systems are subject to unique environmental and social pressures that influence the decision-making of agents and human stakeholders. These pressures create their own underlying behavioural incentives, and changes in these pressures can lead to dramatic shifts in collective behaviour. Characterising the relationships between different environmental conditions and learned behaviours through simulation experiments could help to inform the development of robust methods for incentivising prosocial behaviour under different pressures. Evaluating such methods under systematic variation of environmental conditions could also help to identify the kinds of conditions where methods are susceptible to unintended side effects.

Contributions In this paper, we present ongoing work towards developing EnvCraft, a simulation framework for composing modular environments with parametrised dynamics from reusable components. We propose key features and requirements for EnvCraft to support the study of methods for incentivising prosocial behaviour among self-interested agents under diverse environmental and social pressures.

Organisation The rest of the paper is organised as follows: In Section 2, we detail challenges and experiences from previous work studying prosocial incentives that have informed our proposals for EnvCraft. In Section 3, we describe high-level proposals for building EnvCraft environments. In Section 4, we conclude by discussing directions and objectives for developing EnvCraft as a research tool.

2 Case Study: CustomHarvest

Our aim of developing EnvCraft was motivated by our prior work with CustomHarvest (Figure 1), a simulation environment that we created to study methods for incentivising prosocial behaviour among independent learning agents. In CustomHarvest agents must learn to survive by collecting resources; we designed CustomHarvest as a mixed-motive environment with configurable environment parameters for altering the alignment and conflict between individual and collective incentives. The default parametrisations of CustomHarvest capture social dilemmas, with immediate incentives for individualism, but greater long-term incentives for cooperation. Through parameter selection, long-term individual survival can be made dependent on group survival and resource sharing, i.e. if on average, individual agents use resources faster than they can collect them. By allowing agents to share resources over distances, agents can coordinate as a group to collect resources faster than they use.



Current UUID:
81b81cf4-cac2-4895-bdfc-87ffab177757

Agent Population
pygame_manual: 0
path_finding_agent: 8
path_finding_agent_no_throw: 0

Environment Configuration
grid_shape (Cols): 10
grid_shape (Rows): 10
max_health: 10
max_energy: 4

Fig. 1. Screenshot of recent version of the CustomHarvest environment, including UI elements for configuring simulation runs with different environment parameters. In CustomHarvest agents navigate a grid world and collect berries to survive. Berries are resources that are consumed to restore energy, which agents use energy to move, and health. Agents become helpless if they run out of energy, meaning they stop moving, and start to lose health over time. When agents run out of health, they are removed from the simulation. Agents that become helpless can only survive if another agent shares one of their collected berries with them by throwing it from a distance. Code available https://github.com/dec2g14/custom-harvest.

In Collins et al. [4], we used CustomHarvest to evaluate a novel method for incentivising prosocial behaviour among independent reinforcement learning agents. With this method, agents model a self-imposed "implicit" sense of responsibility for the welfare of others by self-penalising for (A) failing to help agents in need if (B) they were aware that the agent was in need of help (C) they were capable of helping without significant risk to themselves. In this work, we devised CustomHarvest-specific rules for formulating conditions (A-C) in terms of the time and resources needed for survival. In a social dilemma parametrisation, we observed greater sample efficiency in learning to survive cooperatively compared to a population of baseline RL agents [4]. However, experiments were not repeated in different parametrisations in this work. In subsequent experiments, we found certain CustomHarvest parametrisations where our method resulted in poorer collective survival times unexpectedly. With any method for incentivising desirable behaviour, both in simulation and the real world, there is an interplay between incentives arising from the intervention and the underlying pressures of the environment. This can lead to unexpected and unintended consequences if the relationships between environment conditions and behavioural incentives are poorly understood. We highlight this as an important consideration for future work.

A recent version of CustomHarvest (Figure 1) includes a simple interface for exploring the effects of parametrisation on population behaviour for fixed-policy agents. The parametrisation of the dynamics of CustomHarvest is an useful feature for enabling systematic characterisation of the effects of changes in environment dynamics on population behaviour. To better understand these parameter effects, comparisons of RL agent behaviour during training could provide a more robust characterisation of behavioural incentives across different environment configurations. Repeating training over all parameter combinations would be prohibitively expensive using CustomHarvest. However, implementing environments using JAX [3] may provide sufficient improvements in training speed through end-to-end GPU acceleration of both the simulation and model training to enable these experiments.

In ongoing work, we are adapting our implicit responsibility [4] method to formulate conditions (A-C) in terms of empowerment [7], an information-theoretic measure capturing information about an agents potential influence over future states of the environment. Empowerment maximization as an intrinsic motivation can enable agents to learn survival strategies in the absence of extrinsic reward. In very simple environments, such as discrete grid worlds with deterministic or pseudo-random dynamics, the empowerment of an agent in a given state can be computed directly using a simplified version of the Blahut-Arimoto (BA) algorithm [7]. Similar to model-based planning, empowerment computation using the simplified BA algorithm involves sampling all possible successor states from a state transition function for some number of steps into the future. To enable experiments with agents that use BA algorithm or model-based planning methods, CustomHarvest and similar environments would benefit from stateless simulation logic, where state transitions are implemented as pure functions of a state and set of actions. If all state information is contained within the state argument, state transition functions can progress the environment from arbitrary states, supporting planning, and evaluation of trajectories from arbitrary states of interest.

Implementing state transitions as pure functions of state would also support the use of functional JAX [3] operations. Writing simulation and training algorithms in JAX enables end-to-end GPU acceleration, offering large improvements in training speed [6] that may be necessary systematic exploration of environment parameters.

3 EnvCraft Design

This section details high level principles, design features and proposed approaches for developing and implementing EnvCraft. For simplicity, we focus on requirements for the modular composition of discrete grid world environments.

Modular Composition EnvCraft environments should be readily configurable from a specification of reusable modular building blocks. To achieve this, we consider the following abstractions based on the Entity-Component-System

(ECS) architecture. *Entities* are representations of different types of object that interact with an environment. We define entities as configurable sets of *Components*, named data structures used to represent and access variable properties of entities. *Systems* define the state transition dynamics of an environment. To enable modular composition of state transition logic, we define systems as configurable sequences of *Callables*, individual functions for accessing and operating on sets of components. To enable parametrisation of the dynamics of EnvCraft environments by default, all components and callables should be implemented with configuration *Parameters* that are specified when composing new entities and systems. For example, a numerical component would have a parameter for specifying upper and lower bounds, and a callable may require a parameter argument to adjust its operation.

Arbitrary States and Automatic Indexing To enable functionally pure state transitions while preserving environment modularity, EnvCraft should use components implemented as static data structures that contain tuples of unique indices with equal length to their associated data. Rather than treating components as objects that maintain their own internal variables, the indices stored by components can be used to access associated data from a global state array s. For example, a 2D position component representing a coordinates property (x, y) and bounds parameters $(min_x, min_y, max_x, max_y)$ would contain two tuples of indices with lengths 2 and 4. By iterating through indices, starting from zero, each component and parameter for a given environment setup can assigned enumerated indices at instantiation such that the final index gives the dimensions of s needed to represent the environment. This approach enables state transitions to be implemented as configurable sequences of pure functions that take in s and an array of indices as input, access relevant data from s, perform some operation, then return a new state array. EnvCraft should provide an intuitive interface for interacting with sets of environment elements, e.g. by providing methods for indexing a set of components by passing sets names of components and entities.

4 Discussion

A complete implementation of EnvCraft as a software tool is currently an early work-in-progress. In this section, we conclude by discussing future directions and objectives for developing EnvCraft to enable further use cases in prosocial MAS research.

System Characterisation and Robustness Evaluation The parametrised design of EnvCraft environments enables precise exploration of variations in environment pressures in arbitrary environments, readily facilitating experiments for the characterisation of relationships between environmental conditions, population characteristics, emergent behaviours, and collective outcomes. Modular composition allows users to introduce new components and systems to investigate how population dynamics change under different interventions. Future work will focus on extending EnvCraft for evaluating the robustness of prosocial incentive methods under perturbations in environment dynamics, particularly focusing on how heterogeneous agent characteristics interact with environmental pressures and influence the underlying behavioural incentives in an environment.

GPU Acceleration The proposed implementation of EnvCraft using functional state transitions through indexing of arbitrary states supports the use of JAX [3] for GPU acceleration of both environment simulation and agent training. Integrating hardware acceleration is critical for realising the benefits of EnvCraft.

Dynamic Environment Changes Functionality should be added to EnvCraft for dynamic parameter schedules, allowing certain setup parameters to vary according to specified dynamics at run-time. Such features could support work promoting social good in open-ended multi-agent learning environments.

Bibliography

[1] Agapiou, J.P., Vezhnevets, A.S., Duéñez-Guzmán, E.A., Matyas, J., Mao, Y., Sunehag, P., Köster, R., Madhushani, U., Kopparapu, K., Comanescu, R., Strouse, D., Johanson, M.B., Singh, S., Haas, J., Mordatch, I., Mobbs, D., Leibo, J.Z.: Melting pot 2.0 (2023), https://arxiv.org/abs/2211.13746

[2] Ajmeri, N., Guo, H., Murukannaiah, P.K., Singh, M.P.: Elessar: Ethics in norm-aware agents. In: Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). IFAAMAS, Auckland (May 2020). https://doi.org/10.5555/3398761.3398769

[3] Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., Zhang, Q.: JAX: composable transformations of Python+NumPy programs (2018), http://github.com/jax-ml/jax

[4] Collins, D.E., Houghton, C.J., Ajmeri, N.: Fostering multi-agent cooperation through implicit responsibility. In: Proceedings of the 2nd International Workshop on Citizen-Centric Multiagent Systems (CMAS). pp. 1–10. Auckland (2024)

[5] Leibo, J.Z., Zambaldi, V.F., Lanctot, M., Marecki, J., Graepel, T.: Multi-agent reinforcement learning in sequential social dilemmas. CoRR (2017), http://arxiv.org/abs/1702.03037

[6] Rutherford, A., Ellis, B., Gallici, M., Cook, J., Lupu, A., Ingvarsson, G., Willi, T., Khan, A., de Witt, C.S., Souly, A., Bandyopadhyay, S., Samvelyan, M., Jiang, M., Lange, R.T., Whiteson, S., Lacerda, B., Hawes, N., Rocktäschel, T., Lu, C., Foerster, J.N.: Jaxmarl: Multi-agent RL environments in JAX. CoRR (2023), https://doi.org/10.48550/arXiv.2311.10090

[7] Salge, C., Glackin, C., Polani, D.: Empowerment–An Introduction, pp. 67–114. Springer Berlin Heidelberg, Berlin, Heidelberg (2014). https://doi.org/10.1007/978-3-642-53734-9_4

[8] Smit, M., Santos, F.P.: Learning fair cooperation in mixed-motive games with indirect reciprocity. In: Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence. p. 220–228. IJCAI-2024, International Joint Conferences on Artificial Intelligence Organization (Aug 2024). https://doi.org/10.24963/ijcai.2024/25, http://dx.doi.org/10.24963/ijcai.2024/25

[9] Terry, J., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L.S., Dieffendahl, C., Horsch, C., Perez-Vicente, R., et al.: Pettingzoo: Gym for multi-agent reinforcement learning. Advances in Neural Information Processing Systems **34**, 15032–15043 (2021)

[10] Woodgate, J., Marshall, P., Ajmeri, N.: Operationalising rawlsian ethics for fairness in norm-learning agents. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI). pp. 26382–26390. AAAI, Philadelphia (Feb 2025)

3.2 Adaptive Microtolling in Competitive Online Congestion Games via Multiagent Reinforcement Learning

Adaptive Microtolling in Competitive Online Congestion Games via Multiagent Reinforcement Learning

Behrad Koohy[0000−0002−8115−2887], Sebastian Stein[0000−0003−2858−8857], and Enrico Gerding[0000−0001−7200−552X]

University of Southampton, Southampton, United Kingdom
behrad.koohy@soton.ac.uk

Abstract. Efficient urban traffic management remains a critical challenge, yet traditional congestion games fail to capture the dynamic and competitive nature of real-world transportation systems. We introduce the Multi-Market Routing Problem (MMRP), an online and oligopolistic extension that models competition amongst route providers utilising adaptive microtolling strategies to influence driver behaviour and mitigate congestion. We formally define the MMRP, highlighting the computational complexity of solving the MMRP, and use an adapted version of Proximal Policy Optimisation (PPO) to improve update stability in multiagent environments to address this problem in online settings. Our empirical analysis demonstrates that our PPO-based approach not only matches the performance of existing benchmarks but also significantly enhances equity, reduces travel times for users, and increases profitability for providers.

Keywords: Multiagent Reinforcement Learning · Competitive Games · Congestion Games · Adaptive Pricing · Mechanism Design.

1 Introduction

Urban transportation systems are increasingly burdened by congestion, a challenge that significantly impacts economic productivity and quality of life (4). Traditional Congestion Game (CG) (13) models provide valuable insights into individual route choices and their impacts on traffic flow, but neglect the competitive dynamics present in modern, dynamic, transportation networks, where multiple transportation providers compete in an oligopolistic manner. Effective modelling of competition in modern urban transportation networks provides valuable insights into how to maximise the efficiency of existing infrastructure and guide the strategic development of new transportation systems (19).

Traditional CGs are non-cooperative models where individual players select resources (i.e. routes in the context of transportation CGs) and an associated cost is incurred, which escalates with the popularity of that resource. In traffic management, CGs are essential for simulating user behaviours and decision-making in congested environments (22; 8), offering insights into traffic patterns and identifying bottlenecks for optimisation (5; 17; 7; 1; 15; 2). A detailed discussion of existing applications of CGs to congestion management can be found in **(author?)** (11, 12, 20). While these models offer valuable insights into the impact of individual decisions on overall system performance, they are typically static and offline problems which assume perfect information. These simplifications fail to capture the dynamic and competitive interactions inherent in modern urban transportation networks, where multiple route providers continuously compete in real time. This limitation underscores the urgent need for more sophisticated modelling that more closely reflects the dynamic and competitive nature of urban transportation systems.

To overcome these limitations, we propose a novel framework that extends traditional CGs into an online and competitive setting, referred to as the **Multi-Market Routing Problem (MMRP)**. In our framework, transportation networks are modelled as systems where multiple route providers are able to utilise adaptive pricing to influence the behaviour of transportation users in response to fluctuating traffic conditions and competitive pressures. To solve the MMRP in practice, we propose a multiagent reinforcement learning based approach, utilising Proximal Policy Optimisation, to learn adaptive pricing strategies that effectively manage congestion in real time. Our framework bridges the gap between theoretical models and practical traffic management, and our empirical results demonstrate significant improvements in travel times, equity, and profitability, underscoring its potential impact on intelligent transportation systems.

2 Online Multi-Market Routing Problem

We expand the definition of a Congestion Game (19; 14) to the Multi-Market Routing Problem (MMRP) M, where M is a 6-tuple: $M = (R, V, (\Phi_j)_{j \in V}, (\Theta_j)_{j \in V}, (D_i)_{i \in R}, (\Omega_i)_{i \in R})$ The set $R = \{R_0, \ldots, R_i\}$ is the set of available routes; the set $V = \{V_0, \ldots, V_j\}$ is the set of players. For each player $V_j \in V$, Φ_j denotes the strategy space of player V_j and Θ_j denotes the Value of Time (VoT) of player V_j. For each route $R_i \in R$, $D_i : \{0, \ldots, j\} \to \mathbb{R}$ represents the delay function of the route, mapping the number of players selecting a route to a travel time, and Ω_i represents the route cost strategy. For each player $V_j \in V$, $\Phi_j \subseteq 2^R$ defines the strategy space. We define MMRP as an optimisation problem, where the optimal assignment of an instance of MMRP is one in which players incur the lowest total travel time. When this problem is extended to online scenarios, we use the 7-tuple $\mathcal{OM} = (R, V, (\Phi_j)_{j \in V}, (\Theta_j)_{j \in V}, (\Lambda_j)_{j \in V}, (T_i)_{i \in R}, (\Omega_i)_{i \in R})$ where the variables (R, V, Φ) from the offline 6-tuple definition are not changed. In the online definition, Λ_i represents the entry time of a player V_i, and functions D, Ω are changed to become strategies that depend on time t, becoming $D(x, t)$ and $\Omega(x, t)$.

Given the dynamic and real-time decision-making requirements of the online (MMRP), it is essential to understand the computational challenges posed by the offline version, where all players and routes are known in advance. To this end, we establish the computational intractability of the offline MMRP by proving its NP-Hardness through a reduction to the 3-Partition Problem.

3 Proof of NP-Hardness

The theoretical confirmation of the offline MMRP's NP-Hardness serves as a foundation, informing our algorithmic strategies for both offline and online scenarios, and justifies the necessity for the use of heuristic and learning-based approaches such as Reinforcement Learning, which is utilised to address the online problem.

Theorem 1. *There exists a reduction from the strongly NP-Complete 3-Partition Problem to the MMRP*

Proof. To show that the offline MMRP is NP-Hard, we demonstrate a reduction from a known NP-Complete problem, the 3-Partition Problem (3PP), to the offline MMRP.

The 3PP is formally defined as a multiset $A = \{a_1, a_2, \ldots, a_{3m}\}$ of $3m$ positive integers, a bound B such that each integer satisfies $\frac{B}{4} < a_i < \frac{B}{2}$ for all $a_i \in A$ such that the total sum of the integers is $\sum_{i=1}^{3m} a_i = mB$. An instance of the 3PP is valid if A can be partitioned into m disjoint subsets $A_1, A_2, \ldots, A_m$, each containing exactly 3 integers, such that the sum of the integers in each subset is exactly B.

To show a reduction from the 3PP to MMRP, we construct an instance of the MMRP such that finding an optimal assignment with a specific total delay directly corresponds to a valid instance of the 3PP. In the earlier definition, we define the MMRP as an optimisation problem. In this instance of the problem, we define the MMRP as a decision problem where a valid instance exists if the total cost (i.e. the travel time of players) of the instance is equal to or lower than a given L. The optimisation problem can be formed as multiple instances of the decision problem, where if an MMRP instance M is valid for L, it is necessary to check with $L - 1$ until the (M, L) combination is invalid.

For each integer a_i in the 3-Partition instance, we create a corresponding player V_i. Thus, $V = \{V_1, V_2, \ldots, V_{3m}\}$. We create m routes $R = \{R_1, R_2, \ldots, R_m\}$, each representing a potential subset in the 3-Partition. For each player in V, the strategy space $\Phi_i = \{R_1, R_2, \ldots, R_m\}$, allowing them to select any of the m routes. In this instance of the problem, we set all routes to have a cost strategy of $\Omega_i(R_k) = 0$ for all routes R_k. In turn, this negates the impact of all players' value of time Θ_i for $V_i \in V$, and therefore, Θ_i can be set to any value. For each route R_k, the delay function is defined as:

$$D_k(n) = \begin{cases} \sum_{V_i \in S_k} a_i & \text{if } n = 3 \text{ and } \sum_{V_i \in S_k} a_i = B, \\ \infty & \text{otherwise,} \end{cases}$$

where S_k is the set of players assigned to route R_k. It follows from this definition that all routes must have exactly three players assigned, and the sum of S_k must equal B for an instance to be valid.

To align the objectives between this instance of the MMRP and the 3PP, we set $L = mB$. This reduction aims to find an assignment of players to routes such that the total delay is equal to or below L. As L is set to the sum of all integers, it is impossible for there to exist a solution lower than L, so this instance of the MMRP will only return true if there is a valid partitioning for the 3PP.

To show the bidirectional correspondence between the 3PP and the MMRP, we consider the following two instances:

- If the 3PP instance is valid, then it follows that assigning each subset A_k to the corresponding route R_k ensures that:

  $$D_k(S_k) = \sum_{a_i \in A_k} a_i = B.$$

  Leading to a total delay:

  $$\sum_{R_k \in R} D_k(S_k) = mB = L.$$

  This assignment holds in the MMRP as the total delay is below the value of L.
- If there exists an assignment of players to routes in the MMRP such that the total delay is $L = mB$, then:

  $$\sum_{R_k \in R} D_k(S_k) = mB.$$

  As it follows from the definition of $D_k(S_k)$ that $D_k(S_k) = B$, it follows that R_k has exactly three players assigned with the sum of integers equal to B. This assignment would constitute a valid instance of the 3PP for the original set A.

The analysis highlights the NP-Hardness of the offline MMRP, demonstrating the significant computational challenges in optimising routing assignments through pricing strategies, even with complete information. This complexity makes exact optimisation methods impractical for large-scale instances, necessitating the use of heuristics and adaptive learning-based approaches.

The offline formulation of the MMRP is NP-hard[1], rendering exact optimisation methods computationally intractable for large-scale instances and online problems. To this end, we employ Multiagent Reinforcement Learning (MARL), specifically an adapted version of independent Proximal Policy Optimisation (PPO) (16). The use of PPO over existing MARL approaches is deliberate; the inherently competitive nature of route providers suggests that approaches which require inter-agent coordination, such as centralised learning with decentralised execution, are impractical. Consequently, we employed independent PPO, adapted for increased training stability and suitability in our environment,

[1] Proof omitted due to space constraints.

which allows each provider to optimise its pricing strategy. To adapt PPO for use in the MMRP, we employed separate policy and value networks (21), proven effective in multiagent and highly stochastic settings (6), to enhance stability. To mitigate divergence, we normalised rewards by first running random-agent experiments to compute the mean (μ_R) and variance (σ_R) of rewards, and then applied $\tilde{R}_t = \frac{R_t - \mu_R}{\sigma_R}$. Finally, we enabled multiple parallel experiments to replicate the vectorised actor framework used in single-agent PPO (16).

This approach enables each route provider to dynamically adjust tolls in real time, approximating the complex equilibrium behaviour of the system and effectively managing congestion. Our method thus offers a scalable, adaptive solution that bridges the gap between theoretical complexity and practical real-world traffic management. We consider a two-route network ($R = \{R_1, R_2\}$), akin to the parallel two-link models in (18; 9). Each route's delay is defined by the Bureau of Public Roads volume delay function: $D_i(x) = f_0(1 + \alpha(\frac{x}{f_c})^\beta)$, where x is the number of vehicles at time t, f_0 is the free-flow travel time, f_c is the route capacity, and the calibration parameters $\alpha = 0.68$ and $\beta = 2.73$ align the function with real-world data (10). For the values of (f_c, f_0), we set R_1 as $(15, 20)$ and R_2 as $(30, 20)$ for routes 1 and 2 respectively.

We trained our PPO agent for 4×10^7 steps, with each episode lasting 1000 timesteps. In each episode, the number of players was sampled from a uniform distribution $\mathbb{U}(500, 1000)$ to generalise across varied traffic scenarios. Agents share the same architecture, enabling robust performance without environment-specific tuning. The reward function is defined as profit per timestep to reflect a route provider's objective, and the action space consisted of three discrete actions: increase, maintain, or decrease the price by 1.

For our evaluation, we measured the average travel time and profit per vehicle, and employed the Gini coefficient (3) to quantify inequality in travel times across our simulations. A lower Gini coefficient indicates a more equitable distribution of travel times, while a higher value reveals significant disparities . This multi-faceted evaluation framework not only demonstrates the efficiency of our adaptive pricing strategies, but also rigorously assesses their fairness, providing a comprehensive picture of system performance under our proposed solution.

4 Results

Our results (Table 1) demonstrate that our adapted PPO-based approach significantly outperforms a random pricing agent in a two-route synthetic environment. At 500 players, our method achieves an average travel time of 26.66 timesteps compared to 32.50 timesteps for the Random Agent, while consistently maintaining lower Gini coefficients and yielding higher profits (rising from 13.38 at 500 players to 90.41 at 1000 players). Moreover, under infinite capacity conditions, our agents converge towards equilibrium strategies, confirming that our approach effectively captures equilibrium-like behaviour.

These results underscore the potential of our adaptive pricing strategy to transform real-world traffic management, delivering not only reduced congestion,

Table 1. Adaptive Pricing Results for the Online MMRP.

	|V|	**500**	**600**	**650**	**700**	**750**	**800**	**900**	**1000**
MMRP-PPO	Time	**26.66**	**30.85**	**36.24**	**51.58**	**117.1**	**255.65**	**469.12**	**1739.88**
	Gini Coef.	**0.14**	**0.14**	**0.17**	**0.26**	**0.33**	**0.25**	**0.18**	**0.13**
	Profit	13.38	24.65	36.67	47.54	**66.04**	**77.91**	**86.78**	**90.41**
PPO	Time	30.45	35.18	38.34	66.95	108.31	257.13	791.88	1863.87
	Gini Coef.	0.15	0.14	0.17	0.44	0.44	0.33	0.25	0.15
	Profit	34.08	29.95	31.87	33.95	30.16	33.70	39.88	40.94
Random	Time	32.50	34.87	40.14	68.67	151.83	291.61	553.88	2083.49
	Gini Coef.	0.18	0.15	0.18	0.39	0.44	0.37	0.20	0.14
	Profit	**46.13**	**50.8**	**46.64**	**49.61**	47.28	49.94	48.52	46.55

but also a fairer distribution of travel costs. The robust convergence towards equilibrium under infinite capacity further validates our approach, suggesting its applicability in more complex, real-world scenarios.

5 Conclusion

In this study, we introduced the **Multi-Market Routing Problem (MMRP)**, an online, oligopolistic extension of traditional congestion games that models real-world traffic competition through multiple route providers employing adaptive microtolling. We formally defined MMRP and, to overcome its computational complexity, we developed an enhanced **Proximal Policy Optimisation (PPO)** algorithm tailored for competitive multiagent settings. Our evaluations demonstrate that our approach significantly reduces travel times, promotes equity, and increases provider profitability compared to benchmarks. Future work will explore scalability, reduced training costs, advanced techniques such as opponent modelling, and improved explainability to further bridge theory and practical traffic management. Overall, our contributions advance congestion game theory and offer actionable strategies for developing intelligent, adaptive transportation systems.

Acknowledgments. We would like to acknowledge Dr Vahid Yazdanpanah for his feedback on this work. This research is supported by an ICASE studentship from EPSRC and Yunex Traffic, the EPSRC AutoTrust platform grant (EP/R029563/1), and an EPSRC Turing AI Acceleration Fellowship on Citizen-Centric AI Systems (EP/V022067/1).

Bibliography

[1] Brown, P.N., Marden, J.R.: Optimal mechanisms for robust coordination in congestion games. IEEE Transactions on Automatic Control **63**(8), 2437–2448 (2017)

[2] Correa, J., Hoeksma, R., Schröder, M.: Network congestion games are robust to variable demand. Transportation Research Part B: Methodological **119**, 69–78 (2019)

[3] Dorfman, R.: A formula for the gini coefficient. The review of economics and statistics pp. 146–149 (1979)

[4] Dresner, K., Stone, P.: Multiagent traffic management: A reservation-based intersection control mechanism. In: Autonomous Agents and Multiagent Systems, International Joint Conference on. vol. 3, pp. 530–537. Citeseer (2004)

[5] Griesbach, S.M., Hoefer, M., Klimm, M., Koglin, T.: Public signals in network congestion games. In: Proceedings of the 23rd ACM Conference on Economics and Computation. pp. 736–736 (2022)

[6] Huang, S., Dossa, R.F.J., Raffin, A., Kanervisto, A., Wang, W.: The 37 implementation details of proximal policy optimization. In: ICLR Blog Track (2022), https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/, https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/

[7] Khan, Z., Koubaa, A., Benjdira, B., Boulila, W.: A game theory approach for smart traffic management. Computers and Electrical Engineering **110**, 108825 (2023)

[8] Massicot, O., Langbort, C.: Public signals and persuasion for road network congestion games under vagaries. IFAC-PapersOnLine **51**(34), 124–130 (2019)

[9] Milchtaich, I.: Congestion games with player-specific payoff functions. Games and economic behavior **13**(1), 111–124 (1996)

[10] Neuhold, R., Fellendorf, M.: Volume delay functions based on stochastic capacity. Transportation research record **2421**(1), 93–102 (2014)

[11] de Palma, A., Fosgerau, M.: Dynamic and static congestion models: A review. HAL Working Papers (hal-00539166) (2010)

[12] de Palma, A., Lindsey, R.: Traffic congestion pricing methodologies and technologies. Transportation Research Part C: Emerging Technologies **19**(6), 1377–1399 (2011)

[13] Rosenthal, R.W.: A class of games possessing pure-strategy nash equilibria. International Journal of Game Theory **2**(1), 65–67 (Dec 1973). https://doi.org/10.1007/BF01737559, https://doi.org/10.1007/BF01737559

[14] Roughgarden, T., Tardos, É.: Bounding the inefficiency of equilibria in nonatomic congestion games. Games and economic behavior **47**(2), 389–403 (2004)

[15] Scarsini, M., Schröder, M., Tomala, T.: Dynamic atomic congestion games with seasonal flows. Operations Research **66**(2), 327–339 (2018)

[16] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

[17] Swamy, C.: The effectiveness of stackelberg strategies and tolls for network congestion games. ACM Transactions on Algorithms (TALG) **8**(4), 1–19 (2012)

[18] Tavafoghi, H., Teneketzis, D.: Informational incentives for congestion games. In: 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton). pp. 1285–1292. IEEE (2017)

[19] Toșa, C., Mitrea, A., Sato, H., Miwa, T., Morikawa, T.: Economic growth and urban metamorphosis. Journal of Transport and Land Use **11**(1), 273–295 (2018)

[20] Yang, L., Chen, X., Liu, Z.: Impact analysis of congestion pricing scheme in a multimodal transport network. In: CICTP 2019, pp. 3237–3248. ASCE (2019)

[21] Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., Wu, Y.: The surprising effectiveness of ppo in cooperative multi-agent games. Advances in Neural Information Processing Systems **35**, 24611–24624 (2022)

[22] Zhang, J., Lu, J., Cao, J., Huang, W., Guo, J., Wei, Y.: Traffic congestion pricing via network congestion game approach. Discrete & Continuous Dynamical Systems: Series A **41**(7) (2021)

3.3 Combining Human-Centric Modeling and System Optimization in Rural Microtransit

Combining Human-Centric Modeling and System Optimization in Rural Microtransit

Divya Sundaresan[1] [0009−0005−2680−0190], Danushka Edirimanna[2] [0000−0002−5652−161X], Eleni Bardaka[1] [0000−0001−8306−4939], Samitha Samaranayake[2] [0000−0002−5459−3898], and Munindar P. Singh[1] [0000−0003−3599−3893]

[1] North Carolina State University, Raleigh, NC 27606, USA
[2] Cornell University, Ithaca, NY 14850, USA

Abstract. This paper describes ongoing work integrating AI techniques for modeling humans and system optimization in the rural microtransit setting. We consider humans (riders) as central to the system, incorporating their preferences while also accounting for system-wide benefit. Our goal is to develop a microtransit system that maximizes service efficiency, enabling more people to be served with the same resources without increasing costs for riders. We describe the architecture of our system and define our main use cases, as well as our piloting plans and other possible extensions of this work.

Keywords: Civic services · Sociotechnical systems · Public ride sharing

1 Introduction

In many small communities, inhabitants rely on microtransit for their daily transit needs, such as to access employment and healthcare [1]. In these small communities, point-to-point inflexible bus systems are expensive to run as well as underutilized. Hence, microtransit has emerged as a promising solution for connecting suburban and rural populations. Microtransit refers to a shared, technology-enabled public transit system with flexible routing and pickup and dropoff locations that accommodates on-demand trip requests. Rides are booked via an app, but unlike commercial ride-booking systems like Uber and Lyft, microtransit rides are meant to be shared and a nominal fare is charged for usage.

Our goal is to develop, test, and evaluate a smart public microtransit system designed with community-supported solutions for distributing travel demand in an equitable manner [2]. Our contributions are as follows. First, we describe our vision wherein we approach the problem not as a fixed requirement resource allocation problem but as a resource allocation problem where modifications to requirements may be possible at the social tier. We posit that riders may be flexible in their requirements in terms of their pickup and dropoff times if such flexibility would improve some system-wide metric or help another rider. This willingness to adjust one's schedule for another's benefit is an instance of *prosociality*. We use AI-based interventions to understand rider preferences and persuade them to be prosocial [11], and hence serve the commmunity as a whole more effectively. Second, we describe the architecture and use cases of the proposed system that we plan to pilot city-wide in Wilson, NC.

1.1 Motivation: Social Need

In small, disadvantaged communities where the number of zero-vehicle households is high, a large number of inhabitants rely on microtransit to access basic services such as employment and healthcare. The cost of building and maintaining a functional city-wide microtransit system is high, and the lack of funding limits the amount of service that can be provided. This leads to high waiting times, delays, and unserved trip requests.

In the City of Wilson, NC (our partner in this study), microtransit is the only public transit available and is used by a large percentage of the population. In a recent customer survey in Wilson, 47% of the respondents indicated that they use microtransit primarily to travel to and from work, 86% are carless and 57% make less than $25K per year [3]. Wilson's microtransit receives about 18,000 trip requests per month; unfortunately, about 25% of these requests are not served. This poses a daily struggle for inhabitants who depend on microtransit.

1.2 Challenges and Vision

A key motivation for our approach is the synthesis of artificial intelligence with traditional schedule optimization to satisfy rider needs while respecting system constraints. Our goal is to use AI interventions to guide riders toward prosocial behavior. To do so, we must understand rider preferences as well as combine them with system-wide optimization metrics to decide what optimal adjustments are. We posit that in small cities and towns such as Wilson, inhabitants have a strong sense of community. Hence, demand management strategies that motivate prosocial behavior are likely to be more successful than in larger cities and have the potential to significantly improve the day-to-day lives of the transportation disadvantaged groups within these communities.

A sociotechnical system (STS) [8] is a multistakeholder cyberphysical system with a social tier of people and organizations and a technical tier of cyberphysical resources and data. Our contributions are as follows. First, we describe our vision wherein we model the microtransit setting as an STS where its stakeholders (including users and providers, i.e., riders, drivers, and the city transit authority) form the social tier and its cyberphysical resources and data (i.e., vehicles and the associated information technology to request rides) form the technical tier. Since solving the problem at the technical tier is not computationally feasible, we propose that appealing to riders' empathy for other riders and considering rider preferences as flexible (i.e., intervening at the social tier) will make the problem tractable. We hence define our research question as follows:

$\mathbf{RQ_{prosociality}}$: Can framing microtransit optimization as a socially flexible problem—encouraging riders to adopt prosocial behavior and make adjustments—lead to a more efficient system by enabling more requests to be served with the same resources, reducing wait times, and enhancing rider satisfaction?

2 Architecture

We now describe the architecture of the proposed system. We follow conventional architecture design that has been used in previous deployment of similar systems [12]. Figure 1 shows the main modules and flow of the system. First, a rider books a trip on a mobile app (Rider App), where a trip request consists of the origin, destination, and preferred time of pickup. The Backend contains the updated state of the system, such as the realtime position of all the vehicles and trips in progress, as well as scheduled trips, which it passes to the Coordinator along with the incoming request. The Coordinator is a per-city module that considers viable options (where options are alternative time slots during which the incoming request could be scheduled) based on the World Model (which contains the current scheduled trips) and scores them (with the Scorer) based on system benefit and rider preferences. The system benefit of a time slot is calculated by considering both historical demand (a measure of the history-based expected demand for the time slot) and current demand (the scheduled bookings in the time slot as of the current time). Lower demand slots and slots that align with currently scheduled rides have higher system benefit. Rider preferences between time slots are computed by the Rider Agent module, which is designed to model rider behavior by learning their preferences and travel patterns. There is one Rider Agent for each rider. The Rider Agent initializes expected preference over time slots for a rider based on their employment type. Based on repeated interactions, where riders accept specific time slots and indicate their satisfaction with the option, the Rider Agent updates its knowledge of the rider's preferences.

Given a set of optimal alternative time slots, the Optimizer evaluates which of them are feasible. The top three feasible options are provided to the rider, who chooses the one they prefer and indicates their satisfaction through a simple feedback system. This feedback is used by the Rider Agent to learn rider preferences, and the app displays the updated trip with the rider's choice.

We also have a few database components. Demand History contains the history of previous requests and trip demand history, which we use to calculate historical demand. Rider DB contains editable information about each rider such as their home and work address. Rider Trip Logs contains appendable information about trips taken by riders. Rider DB and Rider Trip Logs are used to learn rider preferences since we consider riders' self-specified data, feedback, and travel history to understand them.

We develop the system to accommodate three use cases: on demand, prescheduling, and commuter programs. We describe each use case below.

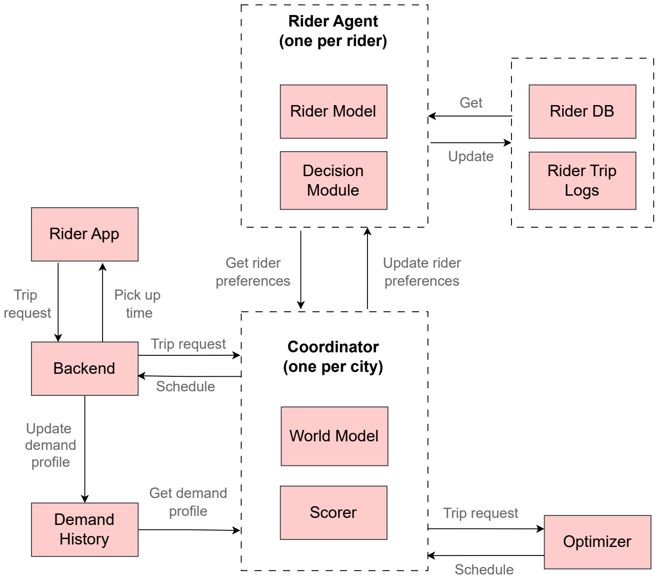


Fig. 1. The main modules of the system.

2.1 Use Case: On Demand

In the on demand use case, riders book trips in real time. When an on demand trip request is received, we aim to serve it immediately (within the next thirty minutes) if there is space in the system to incorporate the trip. Otherwise, the Rider Agent checks whether the ride is flexible, where flexible is defined as a trip that can afford to be pushed to later in the day. Trips to work, the doctor, and back home (since riders should not be left stranded outside) are considered inflexible, while trips to go shopping are considered flexible. If the trip is inflexible, we serve it as soon as possible (during the earliest available time slot). If the trip is flexible, we calculate alternative time slots during which the rider may take this trip that would both align with their preferences and benefit the system. These alternative time slots are calculated by considering the destination hours, the Rider Agent's understanding of rider preferences, and the system benefit, calculated based on historical demand, current demand, and alignment with scheduled trips. Figure 2 shows the sequence of interactions between the components of our system in the on demand use case.

2.2 Use Case: Prescheduling

In the prescheduling use case, we allow riders to request a ride for the next calendar day. After the rider submits their trip request, the Rider Agent analyzes whether the trip is flexible or not, similar to the on demand case. If inflexible, we

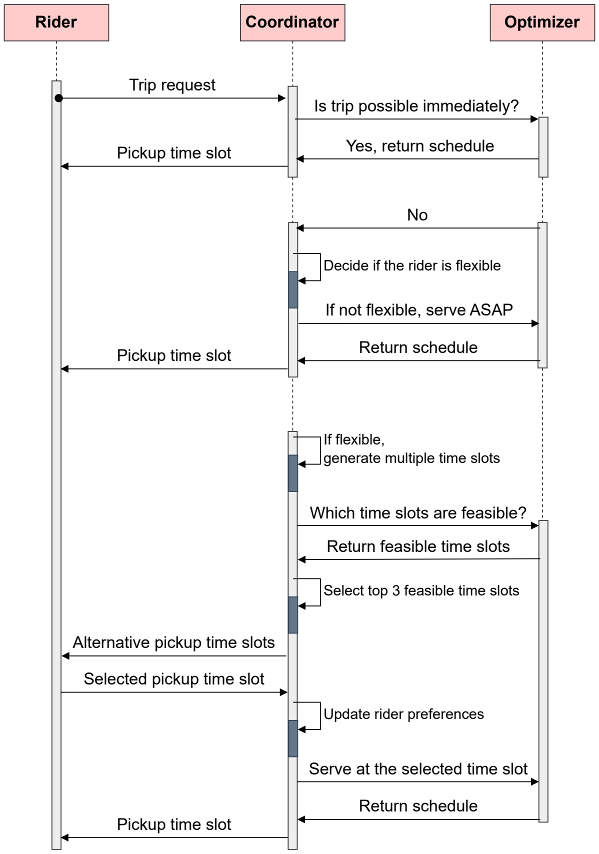

Rider
Coordinator
Optimizer
Trip request
Is trip possible immediately?
Yes, return schedule
Pickup time slot
No
Decide if the rider is flexible
If not flexible, serve ASAP
Return schedule
Pickup time slot
If flexible,
generate multiple time slots
Which time slots are feasible?
Return feasible time slots
Select top 3 feasible time slots
Alternative pickup time slots
Selected pickup time slot
Update rider preferences
Serve at the selected time slot
Return schedule
Pickup time slot

Fig. 2. Sequence of interactions between the rider, the rider agent, and the optimizer in the on demand use case.

allow the rider to specify a 30 minute pickup window. If this window is saturated, we check the windows closest to it until we find an available slot.

If the trip is calculated to be flexible, riders are asked to specify a pickup window which we default to two hours during a low demand time. Riders are prompted to select a larger pickup window if they are flexible, but have the option to decrease it to a 30 minute window. If they select a large window, we schedule them based on system benefit in a time slot within that window. If they choose a small window during a busy time, we request them to shift their trip to another time, similar to the on demand use case. While selecting these alternative time options, we adopt the same approach as in the on demand use case.

2.3 Use Case: Commuter Program

We allow a fixed number of eligible riders to join our subscription-based commuter program. Our optimizer arranges commuters on rides by identifying common desired arrival times and pickup and dropoff locations.

Our contribution is to develop optimization algorithms that enable the integration of commuter programs with our on demand microtransit system. This introduces significant complexity, because some rides need to be scheduled in advance, while also retaining space to service on demand requests.

3 Discussion

We present a conception of a prosocial approach to microtransit by combining human-centric modeling and system optimization. We describe how modeling the microtransit problem in this way can increase system efficiency by increasing the requests served with the same resources, reducing wait times, and increasing rider satisfaction. We describe ongoing work detailing the architecture of the system we plan to pilot in Wilson, NC, and outline our main use cases.

3.1 Piloting Plans

To evaluate our system in a real-world setting, we plan to operate a pilot microtransit service in the City of Wilson, NC, for four months using four microtransit vehicles. Wilson's current microtransit service provides a valuable point of comparison to assess the impact of our AI-based approach. This pilot will allow us to measure key performance indicators such as service rate improvements, rider satisfaction, and the effectiveness of AI-driven interventions in optimizing scheduling. In addition, we aim to measure changes in riders' prosociality: whether they accept more of our prosocial interventions over time, and how satisfied they are with these adjustments. The insights gained will inform further refinements, ensuring that the system is both efficient and equitable before broader deployment.

3.2 Future Work

There is growing interest in developing multi-modal transit systems that integrate mass transit (public transit buses, trams, subways) with microtransit, particularly in low-density demand settings where traditional mass transit is less efficient [4, 10]. Extending our AI-driven microtransit framework to a multimodal setting is a natural extension of this work.

Another direction for future research is to incentivize people to act prosocially by walking further (rather than adjusting their times). In congested urban settings, small location adjustments could greatly benefit the system. Finally, it is necessary to ensure that riders' trust is maintained while building systems that attempt to persuade them to behave differently. An understanding of trust [5, 7], consent [9] and privacy requirements [6] is required to design agents that do not violate riders' autonomy.

Acknowledgments. We acknowledge support from the US National Science Foundation (grant SCC-2325720).

References

1. Bardaka, E., Hajibabai, L., Singh, M.P.: Reimagining ride sharing: Efficient, equitable, sustainable public microtransit. IEEE Internet Computing (IC) **24**(5), 38–44 (Sep 2020). https://doi.org/10.1109/MIC.2020.3018038
2. Bardaka, E., Hentenryck, P.V., Lee, C.C., Mayhorn, C.B., Monast, K., Samaranayake, S., Singh, M.P.: Empathy and AI: Achieving equitable microtransit for underserved communities. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI), Special Track on *AI for Good*. pp. 7179–7187. IJCAI, Jeju, Korea (Aug 2024). https://doi.org/10.24963/ijcai.2024/794
3. Bardaka, E., Hentenryck, P.V., Lee, C.C., Mayhorn, C.B., Monast, K., Samaranayake, S., Singh, M.P.: Empathy and AI: Achieving equitable microtransit for underserved communities. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI), Special Track on *AI for Good*. pp. 7179–7187. IJCAI, Jeju, Korea (Aug 2024). https://doi.org/10.24963/ijcai.2024/794
4. Guan, H., Basciftci, B., Van Hentenryck, P.: Path-based formulations for the design of on-demand multimodal transit systems with adoption awareness. INFORMS Journal on Computing **36**(6), 1459–1480 (2024)
5. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. The Academy of Management Review **20**(3), 709–734 (Jul 1995). https://doi.org/10.5465/amr.1995.9508080335
6. Ogunniye, G., Kokciyan, N.: A survey on understanding and representing privacy requirements in the internet-of-things. Journal of Artificial Intelligence Research **76**, 163–192 (2023). https://doi.org/10.1613/jair.1.14000
7. Singh, A.M., Singh, M.P.: Wasabi: A conceptual model for trustworthy artificial intelligence. IEEE Computer **56**(2), 20–28 (Feb 2023). https://doi.org/10.1109/MC.2022.3212022
8. Singh, M.P.: Norms as a basis for governing sociotechnical systems. ACM Transactions on Intelligent Systems and Technology **5**(1), 1–23 (Jan 2014). https://doi.org/10.1145/2542182.2542203, https://doi.org/10.1145/2542182.2542203
9. Singh, M.P.: Consent as a foundation for responsible autonomy. Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI) **36**(11), 12301–12306 (Feb 2022). https://doi.org/10.1609/aaai.v36i11.21494, blue Sky Track
10. Stiglic, M., Agatz, N., Savelsbergh, M., Gradisar, M.: Enhancing urban mobility: Integrating ride-sharing and public transit. Computers & Operations Research **90**, 12–21 (2018)
11. Sundaresan, D., Watson, A., Bardaka, E., Lee, C.C., Mayhorn, C.B., Singh, M.P.: Prosociality in microtransit. Journal of Artificial Intelligence Research (JAIR) **82**, 77–110 (Jan 2025). https://doi.org/10.1613/jair.1.16777
12. Van Hentenryck, P., Riley, C., Trasatti, A., Guan, H., Santanam, T., Huertas, J.A., Dalmeijer, K., Watkins, K., Drake, J., Baskin, S.: Marta reach: Piloting an on-demand multimodal transit system in atlanta. arXiv preprint arXiv:2308.02681 (2023)

4 Multiagent Learning for Public Decision-Making

4.1 Indirect Reciprocity in Hybrid Human-AI Populations

Indirect Reciprocity in Hybrid Human-AI Populations

Alexandre S. Pires and Fernando P. Santos

Informatics Institute, University of Amsterdam, The Netherlands
{a.m.dasilvapires,f.p.santos}@uva.nl

Abstract. Indirect reciprocity (IR) is a key mechanism to explain cooperation in human populations. With IR, individuals are associated with reputations, which can be used by others when deciding to cooperate or defect: the costs of cooperation can therefore be outweighed by the long-term benefits of keeping a specific reputation. IR has been studied assuming human populations. However, social interactions involve nowadays artificial agents (AAs) such as social bots, conversational agents, or even collaborative robots. It remains unclear how IR dynamics will be affected once artificial agents co-exist with humans. In this project we aim to develop game-theoretical models to investigate the potential effect of AAs in the dynamics of human cooperation. We study settings where artificial agents are potentially subject to the different reputation update rules as the remaining individuals in the population. Furthermore, we consider both settings where reputations are public and setting where reputations are privately held. We show that introducing a small fraction of AAs, with a strategy discriminating based on reputation, increases the cooperation rate in the whole population. Our theoretical work contributes to identify the settings where artificial agents, even with simple hard-coded strategies, can help humans solve social dilemmas of cooperation. At this workshop, we hope to discuss future research avenues where citizens preferences, incentives, and strategic adaptation are considered when designing artificial agents to leverage cooperation in hybrid systems.

Keywords: Social Dilemmas · Evolutionary Game Theory · Hybrid Populations · Indirect Reciprocity.

1 Introduction

Altruistic cooperation requires that individuals spend a cost (c) to provide a benefit (b) to others. When $b > c$, cooperation implies a social dilemma: cooperation is socially desirable yet, given the cost involved, refusing to cooperate is the dominant strategy. Explaining cooperation is a fundamental challenge across disciplines [6] and previous research has identified mechanisms to stabilize it [12]. Among these, indirect reciprocity (**IR**) stands as a primary mechanism to enable cooperation between unrelated individuals [13]. **IR** requires that interactions are observed and individuals assigned reputations [2], which spread through the population (e.g., via gossiping [5]). Simply put, under **IR** cooperating today might contribute to build a reputation that leads others to reciprocate tomorrow [28].

Research in **IR** spans many disciplines. This mechanism is intrinsically related to the evolution of morality, culture and was pointed as a crucial component of a cohesive social structure [2]. To this end, evolutionary game theoretical models have been successfully applied to understand reputation dynamics and their influence in human cooperation [15]. Importantly, however, the viability of **IR** as a mechanism to sustain cooperation in hybrid populations – composed of humans and artificial agents (AAs) – remains unknown. Overcoming this research gap is the key goal of this project [1].

Humans are now increasingly co-existing with AI systems, particularly with socially interactive agents [11]. Examples of these include collaborative robots for navigation [19, 30] or education [24]. It is pressing to understand the impacts that AI systems will have on our collective behavior [1] and our ability to trust and cooperate with AI [4, 16, 26]. Previous works have suggested the role of communication, embodiment [10] and the perception of facing an AA as important aspects of cooperation [9] in hybrid populations. In the context of **IR**, one must identify the differences in how humans and AAs are assessed, and the role of AAs that discriminate in pre-defined ways to opponents' reputations.

In our project, here summarized, we aim to provide a step in addressing two central questions related to **IR** in hybrid populations: **1) What is the impact of artificial agents introduced in a human population interacting under IR? 2) Which social norms promote cooperation in a system composed of humans and artificial agents?** Finally, we ask **3) Can the impacts of artificial agents in indirect reciprocity be generalized to the domain of private reputation systems?**

To answer these questions, we develop theoretical models based on evolutionary game theory (EGT) [28] where a finite population composed of adaptive agents (representing humans) and AAs repeatedly play a *donation game*. In this game, an agent, playing as donor, can cooperate (C), that is, donate, or defect (D) with a receiver. As introduced above, to cooperate means paying a cost c to concede a benefit b, where $b > c > 0$. In our model, agents have reputations that are dynamically updated based on a social norm, i.e., rules that map the action of the donor and the reputation of the receiver to a new reputation for the donor [22]. Furthermore, our model also enables different judgments between adaptive agents and AAs by defining distinct social norms used by observers depending on the nature of the donor, allowing us to consider that humans primarily judge artificial agents by their actions, and not their intentions [8].

We show that introducing a small fraction of AAs whose actions are conditioned on reputations can trigger high levels of cooperation in settings where defection previously prevailed (i.e., low b/c ratio) [17]. Finally, we show that artificial agents can mitigate some of the challenges of **IR** when considering private reputations [18]. Next we summarize these results.

[1] The results reported here are originally published in [17] and [18]

2 Results

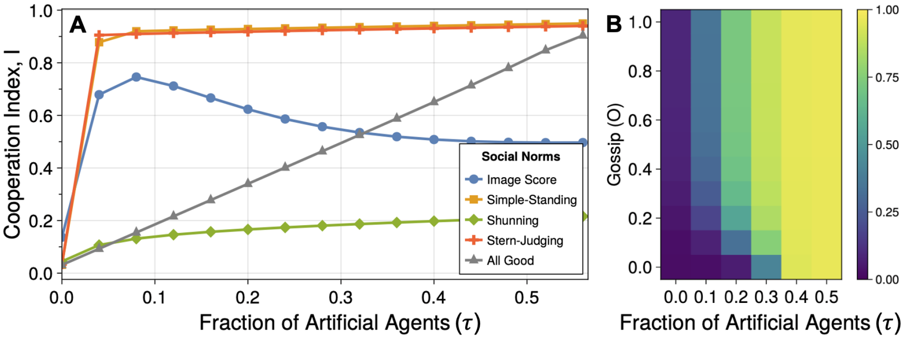


Fig. 1. A: A small fraction of fixed-strategy agents in a population of adaptive agents – simulating artificial agents in a population of humans – can trigger cooperation. This effect is more pronounced for the social norms Image Score (*Cooperation is Good*), Simple Standing (*Cooperation is Good yet Defecting with Bad opponents is justifiable and also Good*) and Stern-Judging (*Cooperating with the Good or Defection with the Bad is Good*) – details in [17]. **B:** Under Simple Standing, the advantages of fixed-strategy agents extend to private reputation systems (low Gossip, O) and settings where individuals judge the AA using simpler norms (i.e., Image-Score) – details in [18]). Parameters: $Z = 100, b = 2, c = 1, e_e = e_a = 0.01, \beta = 1, \gamma = 0.01$.

In Figure 1A, we show how a small fraction of fixed agents (using a *Disc* strategy, see Section 4) can improve cooperation of a majority population of adaptive agents. We focus on a regime where cooperation is challenging to engineer (low benefit to cost ratio, $b/c = 1.2$). However, improvements in cooperation depend on the social norm used. Under Simple Standing and Stern-Judging, where reputations are on average very high, the AA primarily rewards cooperators, leading to high reputations. When using Image Score, we observe a first improvement in cooperation, followed by a decay as the fraction of AAs increases and the population approaches a *Disc* equilibrium, where reputations are neutral. Finally, under Shunning, where reputations are very low due to any interaction with a bad individual leading to a bad reputation, the effect of AAs is less pronounced as they mostly punish every individual. Furthermore, in Figure 1B, we show that interactions with AAs can also improve cooperation under the more strict assumptions that reputations are private (i.e., low gossip, O), adaptive agents do not imitate AAs, and judge the AA using a simplified norm (Image Score).

3 Conclusions

In this project, we investigate cooperation in adaptive populations under the presence of artificial agents (AAs) in the context of indirect reciprocity. The study of **IR** is of particular interest, given the importance of this mechanism in explaining cooperation among unrelated individuals [13] and the possibility that AAs impact **IR** dynamics by acting as donors, receivers, or observers. It is unclear if **IR** will work effectively when artificial systems permeate society.

We developed models to study the impact of artificial agents, implemented with a fixed strategy, integrated in a well-mixed population of adaptive agents. Our results indicate that the effects of such AAs depend primarily on their strategy, as well as the social norm under which humans and AAs are judged. We draw several conclusions: Firstly, the presence of *Disc* Fixed-Strategy agents increases cooperation in previously uncooperative scenarios, under **IS**, **SS** and **SJ** (Figure 1A). This result is of particular importance for **IS**, which is a first-order social norm with low cognitive complexity [21, 23]. Furthermore, if *Disc* AA are evaluated with a positive bias, always being assessed with a good reputation, the previously uncooperative **SH** supports high cooperation levels; cooperation levels under **IS** increase as well (see [17]). Additionally, we highlight that negative biases towards *Disc* AAs result in two opposite forces: an increase of discriminators, which could increase cooperation, and a reduction of good individuals, which typically reduces cooperation. The effect of these agents is thus dependent on the social norm, but also on the benefit of cooperation versus that of defection – we do note that, in general, introducing a low fraction of these agents still augments cooperation. These findings align with the conclusions of other works on cooperation in hybrid populations outside **IR** [20, 3, 7, 25], where low fractions of (pro-social) seeding agents considerably improve cooperation. In addition, we show how AAs are still capable of promoting cooperation in private reputations, where cooperation has been shown to be notably harder to maintain.

The effect of AAs which unconditionally defect is also of great importance, as it highlights a vulnerability of cooperative behavior to uncooperative agents. Our experiments demonstrated that cooperation does not evolve if a low fraction of agents are unconditional defectors. This poses the question of how to develop mechanisms that are resilient against these agents. We also highlight the inefficacy of *AllC* agents, which, due to the dominance of *AllD* among adaptive agents, lead to the exploitation of AAs, which prevents increasing the cooperation levels of the adaptive agents. While a purely theoretical model, these results provide a clear framework and baseline for future human-AI experiments, which can help steer AI development towards a focus on promoting pro-social behavior.

Finally, extracting results from game theoretical models to inform real-world applications and policies requires care [29]. Despite suggesting that discriminating agents can promote cooperation, it is important to note the ethical concerns involved in having autonomous systems dictate what constitutes an acceptable action [27], as well as the fundamental difference in having systems that opt not to cooperate versus ones that actively defect. Our results are constrained to the scope of donation games and indirect reciprocity, discarding eventual risks of over-trusting AI systems. We highlight the need for more thorough human-AI interaction studies [16] in order to bridge the gap between theoretical and experimental results. These works should aim to study how humans judge AAs, both robots and virtual agents, how concrete social norms could be implemented in artificial agents, and also how humans interpret and value judgments made by artificial agents. Furthermore, larger experiments studying how AAs can influence human cooperation could also be conducted in formats similar to our theoretical model. Another direction for future work is to study artificial agents using machine learning as opposed to fixed strategies, allowing for a deeper understanding of human adaptation when facing evolving AAs.

4 Methods

We consider a finite and well-mixed population consisting of Z adaptive individuals, following prior work on **IR** [22]. These agents engage in repeated donation games, where an agent, designated as the donor, can either cooperate, **C**, paying a cost c to offer the other agent, the recipient, a benefit b, where $b > c > 0$, or defect, **D**, where no donation is made, and thus no cost is paid. Other agents observe these interactions and hold a view of every other agent. That is, any agent i considers another agent j either **Good (G)** or **Bad (B)**. We define two regimes: *public reputations*, where all agents agree on the reputation of a focal agent as a consequence of gossip, and *private reputations*, where two individuals must not necessarily agree on the reputation of a focal agent. We vary regimes by considering a number T of gossip rounds happening after each interaction, where a randomly picked observer will adopt the opinion of another observer. At $T = 0$ reputations are private, and at $T \to \infty$ disagreement is null and reputations are public. We interpolate between the two regimes using $T = -log(\epsilon) * O * Z, O \in [0, 1]$, where ϵ is a small threshold value.

The action of each agent is guided by its strategy, which in turn utilizes the reputation assigned to the recipient. A strategy is formally represented as a pair $s = (s_G, s_B)$, where s_G and s_B are the probability of cooperating with an agent perceived as **G** and **B**, respectively. Agents adopt one of three specific strategies at any given moment: *AllC* $(1, 1)$, where cooperation is always selected independently of the reputation of the recipient; *AllD* $(0, 0)$, where the donor always defects; and *Disc* $(1, 0)$, where an individual will only donate to good individuals, and defect against bad individuals. Furthermore, we consider execution errors: with probability e_e, a cooperative action results in defection instead.

We model a hybrid population where human agents interact alongside artificial agents (AAs). In it, AAs can fulfill any of the three roles in the donation game: Donor, Recipient, or Observer. AAs also hold views of others and act based on one of the three strategies. As opposed to adaptive agents, the strategy of an AA is predetermined and remains fixed over time [7, 25]. In addition, we assume that AAs are perfectly coordinated in their assessments. We define τ to be the probability that, for any interaction, a human will instead play an AA. The fixed strategy of all AAs is equal and designated by $s_A \in S = \{AllC, AllD, Disc\}$.

Updates to agents' private reputations are governed by second-order social norms [21]. These norms assess the donor's action (**C** or **D**) in light of the receiver's reputation to determine the donor's new reputation. Each norm is specified by a 4-bit tuple $d = (d_{G,C}, d_{G,D}, d_{B,C}, d_{B,D})$, representing the probability of assigning a good reputation in any of the four possible scenarios. This allows for a total of 16 second-order social norms, of which we focus on four key norms known to sustain cooperation [14]: **Image Score (IS)**, $d = (1, 0, 1, 0)$, where cooperating is always good and defecting is always bad; **Simple Standing (SS)**, $d = (1, 0, 1, 1)$, where only defecting against a good individual is bad; **Shunning (SH)**, $d = (1, 0, 0, 0)$, where only cooperating with a good agent is good; and **Stern Judging (SJ)**, $d = (1, 0, 0, 1)$, where both cooperating with good agents and defecting against bad agents is good, and the remaining is bad. As humans and AAs are judged differently [8], two potentially distinct social norms are applied: d^H, for humans, and d^A, for AAs. We allow for assessment errors, where with a probability e_a the reputation of an agent is incorrectly recalled.

Given these dynamics, we model the adoption of strategies by adaptive agents via a birth-death process [28], where two mechanisms exist: mutations (a probability γ of adopting another available strategy) and social learning. The latter is modeled using the *pairwise comparison rule*, otherwise known as the *Fermi update rule*, where an individual will imitate the strategy of another with a probability that increases with the difference in fitness of the two strategies. We calculate the population dynamics using a Markov chain where each state corresponds to a possible strategy configuration of the adaptive population, and the transition probabilities correspond to the probability of a single agent changing between strategies. The full details of how these dynamics are calculated are presented in [18]. Finally, we measure cooperation through a cooperation index, defined by the average fraction of donations in the population at any time-step.

References

1. Akata, Z., Balliet, D., De Rijke, M., Dignum, F., et al.: A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer **53**(8), 18–28 (Aug 2020)
2. Alexander, R.: The biology of moral systems. Routledge (2017)
3. Anastassacos, N., García, J., Hailes, S., Musolesi, M.: Cooperation and Reputation Dynamics with Reinforcement Learning (Feb 2021), arXiv:2102.07523 [cs]
4. Crandall, J.W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., et al.: Cooperating with machines. Nature Communications **9**(1), 233 (2018)
5. Dores Cruz, T.D., Thielmann, I., Columbus, S., Molho, C., Wu, J., et al.: Gossip and reputation in everyday life. Philosophical Transactions of the Royal Society B **376**(1838), 20200301 (2021)
6. Gintis, H.: Solving the puzzle of prosociality. Rationality and Society **15**(2), 155–187 (2003)
7. Guo, H., Shen, C., Hu, S., Xing, J., Tao, P., Shi, Y., Wang, Z.: Facilitating cooperation in human-agent hybrid populations through autonomous agents. iScience **26**(11) (2023)
8. Hidalgo, C.A., Orghian, D., Canals, J.A., De Almeida, F., Martin, N.: How Humans Judge Machines. MIT Press (2021)
9. Ishowo-Oloko, F., Bonnefon, J.F., Soroye, Z., Crandall, J., Rahwan, I., Rahwan, T.: Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. Nature Machine Intelligence **1**(11), 517–521 (2019)

10. Leite, I., Martinho, C., Paiva, A.: Social robots for long-term interaction: a survey. International Journal of Social Robotics **5**, 291–308 (2013)
11. Lugrin, B., Pelachaud, C., Traum, D. (eds.): The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition, vol. 37. ACM, 1 edn. (2021)
12. Nowak, M.A.: Five rules for the evolution of cooperation. Science **314**(5805), 1560–1563 (2006)
13. Nowak, M.A., Sigmund, K.: Evolution of indirect reciprocity. Nature **437**(7063), 1291–1298 (2005)
14. Ohtsuki, H., Iwasa, Y.: The leading eight: social norms that can maintain cooperation by indirect reciprocity. Journal of theoretical biology **239**(4), 435–444 (2006)
15. Okada, I.: A Review of Theoretical Studies on Indirect Reciprocity. Games **11**(3), 27 (Jul 2020)
16. Paiva, A., Santos, F., Santos, F.: Engineering pro-sociality with autonomous agents. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
17. Pires, A.S., Santos, F.P.: Artificial agents facilitate human cooperation through indirect reciprocity. In: ECAI 2024, pp. 3228–3235. IOS Press (2024)
18. Pires, A.S., Santos, F.P.: Artificial agents mitigate the punishment dilemma of indirect reciprocity. In: AAMAS 2025. IFAAMAS (2025)
19. Rosenthal, S., Biswas, J., Veloso, M.M.: An effective personal mobile robot agent through symbiotic human-robot interaction. In: AAMAS. vol. 10, pp. 915–922 (2010)
20. Santos, F.P., Pacheco, J.M., Paiva, A., Santos, F.C.: Evolution of collective fairness in hybrid populations of humans and agents. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6146–6153 (2019)
21. Santos, F.P., Pacheco, J.M., Santos, F.C.: The complexity of human cooperation under indirect reciprocity. Philosophical Transactions of the Royal Society B: Biological Sciences **376**(1838), 20200291 (Nov 2021)
22. Santos, F.P., Santos, F.C., Pacheco, J.M.: Social norms of cooperation in small-scale societies. PLoS Computational Biology **12**, e1004709 (2016)
23. Santos, F.P., Santos, F.C., Pacheco, J.M.: Social norm complexity and past reputations in the evolution of cooperation. Nature **555**(7695), 242–245 (2018)
24. Serholt, S., Barendregt, W., Vasalou, A., Alves-Oliveira, P., Jones, A., Petisca, S., Paiva, A.: The case of classroom robots: teachers' deliberations on the ethical tensions. AI & Society **32**, 613–631 (2017)
25. Sharma, G., Guo, H., Shen, C., Tanimoto, J.: Small bots, big impact: solving the conundrum of cooperation in optional prisoner's dilemma game through simple strategies. Journal of The Royal Society Interface **20**(204), 20230301 (2023)
26. Shirado, H., Kasahara, S., Christakis, N.A.: Emergence and collapse of reciprocity in semiautomatic driving coordination experiments with humans. PNAS **120**(51), e2307804120 (Dec 2023)
27. Shoham, Y., Powers, R., Grenager, T.: If multi-agent learning is the answer, what is the question? Artificial Intelligence **171**(7), 365–377 (2007)
28. Sigmund, K.: The calculus of selfishness. Princeton University Press (2010)
29. Traulsen, A., Glynatsi, N.E.: The future of theoretical evolutionary game theory. Philosophical Transactions of the Royal Society B **378**(1876), 20210508 (2023)
30. Veloso, M., Biswas, J., Coltin, B., Rosenthal, S.: Cobots: robust symbiotic autonomous mobile service robots. In: Proceedings of the 24th International Conference on Artificial Intelligence. p. 4423–4429. IJCAI'15, AAAI Press (2015)

4.2 Hierarchical Multi-Agent Framework for Dynamic Macroeconomic Modeling Using Large Language Models

Hierarchical Multi-Agent Framework for Dynamic Macroeconomic Modeling Using Large Language Models

Zhixun Chen[1][0009−0007−2617−0744], Zijing Shi[1][0009−0001−4872−9234], Yaodong Yang[2][0000−0001−8132−5613], Meng Fang[3][0000−0001−6745−286X], and Yali Du[4][0000−0001−5683−2621]

[1] University of Technology Sydney, Sydney NSW 2007, Australia
[2] Peking University, Beijing 100871, China
[3] University of Liverpool, Liverpool L69 3BX, UK
[4] King's College London, London WC2R 2LS, UK

Abstract. Large Language Models (LLMs) have demonstrated potential in simulating macroeconomic systems by integrating the agent-based models. Unlike rule-based systems or neural networks with fixed learning patterns, LLM agents capture the heterogeneity of economic actors. However, existing LLM-based simulation environments are generally static, maintaining constant government policies. In this study, we introduce a hierarchical framework that incorporates LLM economic agents and an LLM planner capable of formulating policies in response to evolving economic conditions. Utilizing the proposed framework, we further examine the simulated system's resilience to economic shocks by analyzing how economic agents respond to unforeseen events and how the planner adapts to mitigate these challenges. Our results indicate that the proposed framework improves the stability of the economic system and captures more dynamic macroeconomic phenomena, offering a precise and versatile simulation platform for studying real-world economic dynamics.

Keywords: Agent-Based Model · Large Language Model· Macroeconomic Modeling

1 Introduction

The complexity of modern economies has prompted researchers to explore methods for simulating macroeconomic systems, with a particular focus on Agent-Based Models (ABMs) [22, 7]. Early models, relying on rule-based systems or neural networks, struggled to capture the behavioral heterogeneity of real economies [3, 10]. The introduction of neural networks has improved the flexibility and intelligence of modern ABM models by integrating deep learning methods such as reinforcement learning [23, 29].

However, generalization and robustness across different environments remain challenging. Recent advancements in Large Language Models (LLMs) have demonstrated advanced abilities such as reasoning and decision-making [12, 4], enabling them to simulate complex economic activities like trade and resource allocation [21, 13, 25, 27]. EconAgent, which employs LLM agents for macroeconomic simulations, offers a more nuanced representation of economic agents but still treats agent interactions statically and overlooks dynamic government policies and economic shocks [17].

We propose a hierarchical, dynamic multi-agent framework that incorporates LLM agents to simulate economic policy planning and shock resilience. Our framework simulates adaptive agent behavior and evaluates the system's response to economic shocks by enabling LLM agents to adjust policies such as tax rates and inflation targets, thereby capturing interactions between heterogeneous agents and policy planners. Our experiments demonstrate that both LLM planners and agents can detect shocks, leading to swift recovery and enhanced system stability.

2 Method

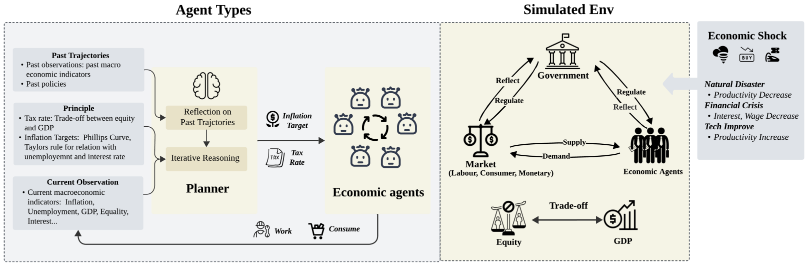

Agent Types
Past Trajectories
• Past observations: past macro economic indicators
• Past policies
Principle
• Tax rate: Trade-off between equity and GDP
• Inflation Targets: Phillips Curve, Taylors rule for relation with unemployemnt and interest rate
Current Observation
• Current macroeconomic indicators: Inflation, Unemployment, GDP, Equality, Interest...
Reflection on Past Trajctories
Iterative Reasoning
Planner
Inflation Target
TAX
Tax Rate
Economic agents
Work
Consume
Simulated Env
Government
Reflect
Regulate
Regulate
Reflect
Market
(Labour, Consumer, Monetary)
Supply
Demand
Economic Agents
Trade-off
Equity
GDP
Economic Shock
BUY
Natural Disaster
• Productivity Decrease
Financial Crisis
• Interest, Wage Decrease
Tech Improve
• Productivity Increase

Fig. 1. An overview of the proposed framework. The planner monitors macroeconomic indicators and analyzes them using the principle module's guidance. It also considers past data to offer suggestions for future economic policy-making. Based on these insights, the planner formulates policies, such as inflation target and tax rates for the next year, aiming to balance equality and GDP within the simulation. These policies influence the behavior of economic agents, influencing their consumption choices and work propensity. Additionally, unexpected economic shocks may arise, impacting the labor, consumer and momentary markets. Both the planner and agents adjust in response, restoring the stability of the simulated environment.

2.1 Hierarchical Multi-Agent Framework

Building on the work of EconAgent [17], we introduce a macroeconomic simulation framework that employs agent-based modeling to capture complex economic interactions, as illustrated in Figure 1. The system consists of a planner and multiple heterogeneous economic agents operating on different timescales. Our framework extends previous efforts by incorporating dynamic planner decision-making and economic shocks to better mirror real-world conditions.

The planner optimizes macroeconomic variables—such as tax rates and inflation targets—on an annual basis, while the economic agents adjust their behavior monthly according to individual preferences and incentives. The planner P observes o_t based on: (1) macroeconomic indicators—such as unemployment rate, inflation rate, GDP growth, average wage, and economic equality—for the past L years, and (2) historical government policies over the past L years. Based on these observations, the planner sets tax rates for each income bracket $(\tau_1, \ldots, \tau_B)$, constrained between τ_{low} and τ_{high}, and determines the target inflation rate π_t for the coming year. For example, when the planner adjusts income tax rates, it affects the post-tax income that agents receive, which in turn alters their utility functions. This multi-agent learning problem, which naturally emerges in various economic and machine learning scenarios [9, 6], resembles a Stackelberg game [24], where the planner, acting as a leader, optimizes long-term economic outcomes while individual agents, as followers, make strategic decisions to maximize their own utilities within the constraints of these policies.

2.2 Grounding LLM as Planner

To effectively simulate the planner's role using an LLM, we incorporate reflective and iterative reasoning processes. The Principle and Observation Module provides macroeconomic guidelines for adjusting tax rates and setting inflation targets based on economic growth, inequality, and stability. Following Laffer curve theory [15], the planner optimizes tax rates to maximize redistribution without hindering economic activity. Additionally, the Taylor Rule [26, 5] and the Phillips Curve [18] guide the planner in balancing inflation and unemployment to enhance social welfare. The Reflection Module reviews historical trajectories by retrieving L prior action-observation pairs, facilitating continuous policy improvement. By analyzing past trajectories, fundamental economic principles, and current observations, the LLM-based planner iteratively refines its decision-making, ensuring adaptive and robust policy formulation.

3 Experiments

Our experiments explore how a planner can control key macroeconomic indicators—such as GDP growth, unemployment, inflation, and equality—through the implementation of tax policies and inflation targets. We set the number of agents to 50. Figure 2 illustrates the system's economic situation during a natural disaster, where the productivity factor A drops sharply, triggering price inflation and reducing GDP. In our method, the unemployment rate spikes to 20% and societal equality falls below 50%. As productivity recovers after five years, inflation stabilizes around 1%, unemployment decreases to 10%, and equality rises above 50%, indicating economic recovery.

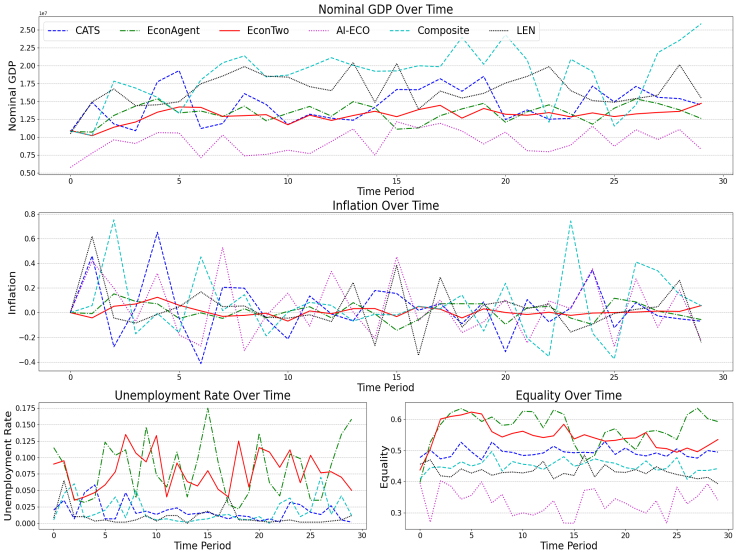


Fig. 2. Variation of annual macroeconomic indicators under an economic shock caused by a natural disaster. The shock occurs in the eighth year of the simulation.

In contrast, the EconAgent environment maintains high societal equality immediately after the shock. However, this hinders economic recovery post-disaster [14]. Rule-based models recover GDP better but show equality levels below 50% (in some cases below 40%), suggesting uneven recovery benefits. Additionally, these models experience extreme inflation fluctuations between -60% and 60%. AI-ECO exhibits 35% equality and 50% unemployment, resulting in consistently low GDP.

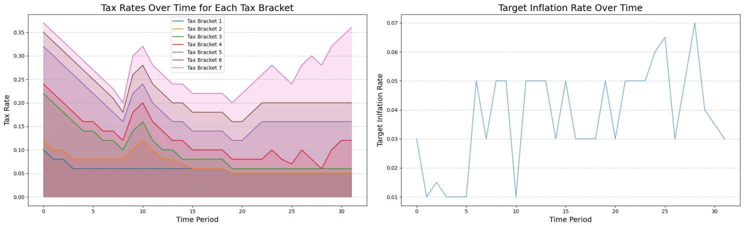


Fig. 3. Planner policies under the economic shock: natural disaster. The planner adjusts economic policies in response to the prevailing conditions of the environment.

We then analyze how agents and the planner adapt to economic conditions. Figure 3 depicts the planner's adjustments to tax rates and inflation targets during a natural disaster, aligning with macroeconomic fluctuations shown in Figure 2. Typically, the planner lowers tax rates to stimulate consumption and GDP growth when the natural disaster occurs. However, if social inequality surpasses a critical threshold, tax rates—especially for high-income earners—are increased to promote equity. Inflation targets are generally set between 3% and 5% to balance consumption and unemployment through moderate interest rates. In response to significant macroeconomic shifts, the planner adopts more aggressive tax and inflation measures to restore stability. For example, in the 10th year, stagflation emerges, characterized by high inflation and unemployment. Since reducing inflation requires higher taxes while lowering unemployment necessitates tax cuts, the planner resolves this conflict by imposing steeper tax increases on higher-income groups while applying smaller adjustments to lower-income brackets. The findings suggest that the planner effectively tailors policies to evolving economic conditions, guiding agents to align with its objectives. This hierarchical approach enhances system sustainability, stability, and resilience against economic shocks.

4 Conclusion

In conclusion, this work advances macroeconomic simulation with the hierarchical framework of economic agents and a dynamic policy planner. Unlike static models, our planner adapts to evolving conditions, aligning better with real-world complexities. By utilizing the principle and reflection modules, it effectively handles economic shocks and enhances social welfare, resulting in a resilient system. The study shows that our method captures intricate economic behaviors, making it a valuable platform for exploring macroeconomic policies. This work highlights the potential of LLMs in simulating complex economic systems, opening new pathways for analyzing responsive policymaking and economic phenomena.

References

1. Aghion, P., Jones, B.F., Jones, C.I.: Artificial intelligence and economic growth, vol. 23928. National Bureau of Economic Research Cambridge, MA (2017)
2. Botzen, W.W., Deschenes, O., Sanders, M.: The economic impacts of natural disasters: A review of models and empirical studies. Review of Environmental Economics and Policy (2019)
3. Chen, S.H., Chang, C.L., Du, Y.R.: Agent-based economic models and econometrics. The Knowledge Engineering Review **27**(2), 187–219 (2012)
4. Chen, X.H., Wang, Z., Du, Y., Jiang, S., Fang, M., Yu, Y., Wang, J.: Policy learning from tutorial books via understanding, rehearsing and introspecting. Advances in Neural Information Processing Systems **37**, 18940–18987 (2025)
5. Dawid, H., Gatti, D.D.: Agent-based macroeconomics. Handbook of computational economics **4**, 63–156 (2018)
6. Du, Y., Leibo, J.Z., Islam, U., Willis, R., Sunehag, P.: A review of cooperation in multi-agent learning. arXiv preprint arXiv:2312.05162 (2023)
7. Farmer, J.D., Foley, D.: The economy needs agent-based modelling. Nature **460**(7256), 685–686 (2009)

8. Gatti, D.D., Desiderio, S., Gaffeo, E., Cirillo, P., Gallegati, M.: Macroeconomics from the Bottom-up, vol. 1. Springer Science & Business Media (2011)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014)
10. van der Hoog, S.: Deep learning in (and of) agent-based models: A prospectus. arXiv preprint arXiv:1706.06302 (2017)
11. Jermann, U., Quadrini, V.: Macroeconomic effects of financial shocks. American Economic Review **102**(1), 238–271 (2012)
12. Jin, X., Wang, Z., Du, Y., Fang, M., Zhang, H., Wang, J.: Learning to discuss strategically: A case study on one night ultimate werewolf. Advances in Neural Information Processing Systems **37**, 77060–77097 (2025)
13. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
14. Kuznets, S.: Economic growth and income inequality. In: The gap between rich and poor, pp. 25–37. Routledge (2019)
15. Laffer, A.B.: The laffer curve: Past, present, and future. Backgrounder **1765**(1), 1–16 (2004)
16. Lengnick, M.: Agent-based macroeconomics: A baseline model. Journal of Economic Behavior & Organization **86**, 102–120 (2013)
17. Li, N., Gao, C., Li, M., Li, Y., Liao, Q.: Econagent: large language model-empowered agents for simulating macroeconomic activities. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15523–15536 (2024)
18. Phelps, E.S.: Phillips curves, expectations of inflation and optimal unemployment over time. Economica pp. 254–281 (1967)
19. Ramey, V.A.: Macroeconomic shocks and their propagation. Handbook of macroeconomics **2**, 71–162 (2016)
20. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
21. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems **36** (2024)
22. Tesfatsion, L.: Agent-based computational economics: A constructive approach to economic theory. Handbook of computational economics **2**, 831–880 (2006)
23. Trott, A., Srinivasa, S., van der Wal, D., Haneuse, S., Zheng, S.: Building a foundation for data-driven, interpretable, and robust policy design using the ai economist. arXiv preprint arXiv:2108.02904 (2021)
24. Von Stackelberg, H.: Market structure and equilibrium. Springer Science & Business Media (2010)
25. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
26. Wolf, S., Fürst, S., Mandel, A., Lass, W., Lincke, D., Pablo-Marti, F., Jaeger, C.: A multi-agent model of several economic regions. Environmental modelling & software **44**, 25–43 (2013)
27. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems **36** (2024)

28. Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D.C., Socher, R.: The ai economist: Improving equality and productivity with ai-driven tax policies (2020), https://arxiv.org/abs/2004.13332
29. Zheng, S., Trott, A., Srinivasa, S., Parkes, D.C., Socher, R.: The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. Science advances **8**(18), eabk2607 (2022)

A Appendix

A.1 Planner Decisions

Planner policies, particularly those involving tax rates and inflation targets, are critical for guiding economic development and ensuring social welfare. These policies significantly affect the decision-making of household agents. In our framework, the government adjusts tax rates and inflation targets on an annual basis. Specifically, the government planner P observes o_t:

- Macroeconomic indicators I_m over the past L years, including unemployment rate, inflation rate, GDP growth, average wage of agents, and economic equality[5]. The observed indicators for the years are represented as $(I_1, ..., I_y, ..., I_L)$.
- Historical government policies over the past L years, allowing the planner to reflect on the effectiveness of previous policy adjustments.

Based on these observations, the government determines the following actions a_p:

- The tax rate for each income bracket $(\tau_1, ..., \tau_k, ..., \tau_B)$, where the tax rate in bracket k is constrained between τ_{low} and τ_{high}, defining the proportion of income that agents in that bracket must contribute in taxes.
- The target inflation rate π_t, representing the desired inflation level for the coming year.

Taxes are collected from all agents' incomes for that year. The progressive tax for agent i's income z_i is computed as follows:

$$T(z_i) = \sum_{k=1}^{B} \tau_k((b_{k+1} - b_k)\mathbf{1}[z_i > b_{k+1}] + (z_i - b_k)\mathbf{1}[b_k < z_i \leq b_{k+1}]) \qquad (1)$$

where $\mathbf{1}[\cdot]$ is the indicator function. The tax brackets follow the 2018 U.S. Federal tax schedule [28]. Following [17, 28], tax revenue is redistributed evenly among all agents. Therefore, the post-tax income of agent i is:

$$\hat{z}_i = z_i - T(z_i) + z_r = z_i - T(z_i) + \frac{1}{N}\sum_{j=1}^{N} T(z_j) \qquad (2)$$

where z_r represents redistribution. Agent savings are updated accordingly:

$$s_i \leftarrow s_i + \hat{z}_i \qquad (3)$$

The target inflation rate guides changes in interest rates. We use the widely accepted Taylor rule to determine interest rates [26]:

$$r = \max(r_n + \pi_t + \alpha_\pi(\pi - \pi_t) + \alpha_u(u_n - u), 0) \qquad (4)$$

[5] Economic equality is measured using the Gini coefficient. Further details are provided in appendix A.6

where r_n and u_n represent the natural interest rate and unemployment rate, respectively. π_t is the target inflation rate, π is the actual annual inflation rate, and u is the annual unemployment rate. The coefficients α_π and α_u are adjustable parameters used to regulate the effects of inflation and unemployment, respectively.

A.2 Economic Shocks

Introducing economic shocks into macroeconomic simulations is essential for accurately modeling real-world economic dynamics. These shocks represent sudden, often unexpected events that significantly alter the course of an economy, such as technological breakthroughs, financial crisis, or natural disaster. Including such shocks allows simulations to reflect the volatility and complexity of the real world, providing insights into how agents and planner adjust to rapid changes in the economic environment [8, 5]. Previous studies have shown that incorporating economic shocks enhances the realism of simulations, allowing for the exploration of how different shocks propagate through an economy and how policies are adapted in response [26, 2, 19, 11].

Economic shocks can be introduced in various global forms. One prominent example is a technological advancement, which leads to an increase in universal productivity. Such a shock would affect all agents by improving production efficiency and raising output [1]. This could be modeled as a positive productivity shock that increases the productivity factor A across the economy:

$$A' = A \times (1 + S_t), \quad S_t \sim U(a_1, b_1) \tag{5}$$

where S_t represents the size of the technological shock, drawn from a uniform distribution between a and b. This shock would have widespread effects, raising wages, lowering the price of goods due to increased supply, and boosting overall consumption.

Another example of a shock is natural disasters or pandemics, which can reduce labor supply or cause significant disruptions to production [2]. A shock to labor supply could reduce the total hours worked across all agents, thereby reducing the productivity factor A. The shock will lead to lower output and a reduced supply of goods:

$$A' = A \times (1 - S_t), \quad w_i' = w_i \times (1 - W_s \cdot h(w_i)) \quad S_t, W_s \sim U(a_1, b_1) \tag{6}$$

where A' represents the reduced productivity due to decreased labor supply and S_t models the shock to the productivity. The equation for w_i' models the impact on wages, where w_i represents the initial wage of agent i, W_s represents the size of the natural disaster that reduces average wage, and $h(w_i)$ is a function that decreases with increasing w_i. This means that lower-income agents face larger proportional reductions, reflecting their vulnerability to natural disasters, whereas higher-income agents are affected to a lesser extent. As a result, this dynamic leads to greater inequality and highlights the uneven economic impacts experienced during such shocks.

Lastly, a financial crisis such as a stock crash can be modeled as a shock to the monetary market of the environment, where agents experience sudden losses in wealth or savings, leading to a sharp contraction in consumption [11]. This scenario can be represented as a reduction in agents' savings and wages:

$$\begin{aligned} s_i' = s_i \times (1 - S_w), \quad w_i' = w_i \times (1 - W_s \cdot f(w_i)), \quad A' = A \times (1 - S_t) \\ S_w, W_s, S_t \sim U(a_2, b_2), \quad r = 0 \end{aligned} \tag{7}$$

where S_w represents the size of the financial shock that reduces savings. The function $f(w_i)$ is defined such that agents with higher initial wages experience proportionally greater reductions. This is based on the notion that higher-income sectors are often more exposed to economic fluctuations, particularly during a stock market crash, which can severely impact executive compensation and high-paying sectors. The productivity A of the society is also reduced to model the stock crash situation because a sharp decline in stock market value can lead to reduced investments and widespread layoffs. In addition, the interest rate r of the environment is force to 0 to mimic real-world situation. As a result, this differential impact forces agents to cut back on their consumption according to their wage levels, potentially leading to a prolonged economic downturn.

In addition, a recovery policy is implemented if the productivity factor A' is less than the original A.

$$A(t) = \max\left(A, A' + r_n \cdot \max(0, t - t_0 - n)\right) \tag{8}$$

where t_0 is the time when the shock occurs, n represents the number of years after which recovery begins, and r_n is the annual recovery rate. This representation ensures that after an initial lag of n years, productivity starts to recover linearly at rate r until it reaches the original value A, after which it stabilizes. This equation aims to simulate the restoration of productivity gradually, mitigating the long-term negative effects of the shock and supporting economic recovery. During the economic shock, additional sentences will be added into the input for both the agents and the planner, enabling them to better perceive the current economic conditions. Once productivity A returns to its original level or $t - t_0 > 5n$, the economic shock is considered to have ended, and the extra sentences will be removed from the input.

A.3 Agent Decisions

In the framework, agent i make decisions monthly on work and consumption:

- Agents decide whether to work, denoted by $l_i \sim \text{Bernoulli}(p_{wi})$, where p_{wi} represents the agent's work propensity. If an agent chooses to work ($l_i = 1$), they receive a monthly wage as income, which differs across agents. Initially, each agent has an hourly wage w_i, drawn from a Pareto distribution [29].

The monthly wage v_i is then calculated by multiplying 168 hours (assuming 21 working days at 8 hours per day [16]). If the agent decides not to work ($l_i = 0$), their income for that month is zero.

- The consumption propensity p_{ci} reflects the proportion of an agent's total wealth (current savings and income in the month) that they intend to allocate for purchasing essential goods.

Modeling the varied decisions of heterogeneous agents is crucial for reflecting macroeconomic dynamics. The decision-making process for each agent is shaped by several economic factors, including anticipated income and tax obligations.

A.4 Productivity and Consumption

By incorporating agent decisions and government taxation, the dynamics of labor and consumption markets are simulated based on economic principles. Agents who work contribute 168 hours of productivity per month, leading to the production of essential goods. The total inventory of goods, G, is updated as:

$$G \leftarrow G + S = G + \sum_{j=1}^{N} l_j \times 168 \times A \tag{9}$$

where S represents the volume of goods produced from the labor supply, and A is the productivity factor. For consumption, the total demand for goods is expressed as:

$$D = \sum_{j=1}^{N} d_j = \sum_{j=1}^{N} \frac{p_{cj} s_j}{P} \tag{10}$$

where d_j is the demand of agent j, p_{cj} is the consumption propensity, s_j is the current savings, and P is the price of essential goods. Both labor and consumption markets update based on the imbalance relation between supply and demand. The imbalance is defined as:

$$\bar{\phi} = \frac{D - G}{\max(D, G)} \tag{11}$$

When there is a shortage of essential goods, meaning demand exceeds supply, wages should increase to encourage production. As labor costs rise, the prices will be updated correspondingly. The hourly wage is updated as:

$$w_i \leftarrow w_i(1 + \phi_i), \quad \phi_i \sim \operatorname{sign}(\bar{\phi}) U(0, \alpha_w |\bar{\phi}|) \tag{12}$$

The price of goods is adjusted similarly:

$$P \leftarrow P(1 + \phi_P), \quad \phi_P \sim \operatorname{sign}(\bar{\phi}) U(0, \alpha_P |\bar{\phi}|) \tag{13}$$

where α_w and α_P are the maximum rates of wage and price adjustments, respectively. The dynamics of consumption are also modeled. Specifically, an agent j

is randomly chosen to consume goods, with actual consumption limited by the available inventory:

$$\hat{d}_j = \min(d_j, G), \quad \hat{c}_j = \hat{d}_j \times P \tag{14}$$

This indicates that demand is met only if there is sufficient supply. The total goods inventory is reduced accordingly:

$$G \leftarrow G - \hat{d}_j \tag{15}$$

This process repeats until each agent has consumed goods once.

A.5 Financial Market

According to the change of interest rate, the savings of each agent increase based on the equation:

$$s_i \leftarrow s_i \times (1 + r) \tag{16}$$

The inflation and unemployment rate are defined as:

$$\pi = \frac{\bar{P}_n - \bar{P}_{n-1}}{\bar{P}_{n-1}}, \quad u = \frac{\sum_{m=1}^{12} \sum_{j=1}^{N} (1 - l_j)}{12N} \tag{17}$$

A.6 Equality

The equality of the society is measured by equality index, which is calculated by:

$$\text{eq}(x^c) = 1 - \frac{\text{gini}(x^c) N}{N - 1} \tag{18}$$

where $x^c = (x_1^c, x_2^c, \ldots, x_i^c, \ldots, x_N^c)$ is the vector of wealth of all N agents after taxation and redistribution. The wealth of one agent means the wage of current month and previous saving: $x_i^c = w_i + s_i$. $\text{gini}(x^c)$ represents the Gini coefficient, a standard measure of inequality, with values ranging from 0 (perfect equality) to 1 (perfect inequality). The Gini coefficient is calculated as:

$$\text{gini}(x^c) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |x_i^c - x_j^c|}{2N \sum_{i=1}^{N} x_i^c} \tag{19}$$

This formula computes the relative differences in wealth between all agents. If all agents have the same wealth, the Gini coefficient is zero, indicating maximum equality. Conversely, a high Gini coefficient indicates greater inequality, with one agent controlling most of the wealth.

A.7 Economic Shock

For the $f(w_i)$ and $h(w_i)$, the general form equation is:

$$g(w_i, \sigma) = 0.8 + 0.2 \times \left(\frac{w_i}{w_{max}} \right) \cdot \sigma \tag{20}$$

Specifically, $f(w_i)$ equals to $g(w_i)$ when $\sigma = 1$, $h(w_i)$ equals to $g(w_i)$ when $\sigma = -1$.

A.8 Baselines

LEN The LEN [16] model is an Agent-Based Model (ABM) that simulates macroeconomic interactions between households and firms. Households decide labor supply and consumption, while firms adjust prices and wages according to economic conditions. In this model, consumption decisions are memory-based, meaning they depend on both current income and past accumulated savings. The consumption propensity is expressed as:

$$p_c^i = \left(\frac{P}{s_i + z_i}\right)^{\beta} \tag{21}$$

where p_c^i is the consumption propensity, P is the price of goods, s_i is the current savings of household i, z_i is the current income, and $\beta \in [0, 1]$ is a parameter.

CATS The CATS [8] model simulates an economy with households, firms, and banks. Unlike LEN, consumption decisions here are non-memory-based, relying only on current income. Households aim to maintain a desired ratio between savings and income. The consumption propensity is given by:

$$\frac{s_i}{z_i} = (1+r)\left(\frac{s_i + (1-c)z_i}{z_i}\right) = h \quad p_i^c = \frac{cz_i}{s_i + z_i} \tag{22}$$

where s_i is the savings, z_i is the income, r is the interest rate, h is the desired savings-to-income ratio, and c is the proportion of income consumed.

For the work rule of these two baselines, we follow EconAgent [17] setup. The equation for work propensity is:

$$p_{wi} = \left(\frac{w_i}{s_i(1+r)}\right)^{\gamma} \tag{23}$$

Composite The composite baseline is a hybrid model where an agent's action is randomly selected from either the CATS or LEN [17]. This allows the agent to alternate between decision-making frameworks based on probabilistic rules. The agent's action, a_i, is determined using a Bernoulli distribution, where the probability of choosing the CATS or LEN model is equal:

$$X \sim \text{Bernoulli}(0.5) \quad a_i = \begin{cases} a_{CATS}, \text{if } X = 0 \\ a_{LEN}, \text{if } X = 1 \end{cases} \tag{24}$$

Saez Formula: The Saez tax formula optimizes tax rates based on the income distribution and the elasticity of income with respect to tax rates. Tax elasticity $e(z)$ measures how sensitive an individual's income is to changes in the tax rate. It is defined as:

$$e(z) = \frac{dz/z}{d\left(1-\tau(z)\right)/(1-\tau(z))},$$

where z is the pre-tax income and $\tau(z)$ is the marginal tax rate. Higher elasticity implies that a small increase in the tax rate significantly reduces the income earned.

We follow the setup of AI-Economist to use Multi-period Saez Formula here. In a multi-period setting, the tax elasticity $\tilde{e}$ is estimated at the start of each tax period, where a buffer $D = \{(z_i^\alpha, \tau_i^\alpha)\}_\alpha$, a set of observed income and tax rate pair is used for estimation.

Then the income is modeled as:

$$z_t = z_0 \cdot (1 - \tau_t)^{\tilde{e}} \tag{25}$$

where z_0 is the hypothetical income without taxation. This can be rewritten as:

$$\log(z_t) = \tilde{e} \cdot \log(1 - \tau_t) + \log(z_0) \tag{26}$$

We estimate $\tilde{e}$ using ordinary least-squares regression on this equation, utilizing the most recent data from several tax periods. This allows stable estimation of the average elasticity $\tilde{e}$, ensuring accurate adaptation of the tax rates in each period. In the experiment, the buffer size is 30,000 with the most recent incomes and tax rates observed during rollout episodes.

Rule-based Inflation Target We apply a rule-based baseline similar to the Taylor rule [26, 5, 17] for adjusting the inflation target rate based on deviations from desired macroeconomic variables such as inflation, unemployment, and GDP growth. The formula is:

$$\pi_{t+1} = \max\left(0, \min\left(0.08, \pi_t + (\beta_\pi \cdot (\pi - \pi_t)) + (\beta_u \cdot (u - u^n)) + (\beta_y \cdot (Y_g - Y_g^n))\right)\right) \tag{27}$$

where the coefficients β_π, β_u, and β_y determine the weight of each macroeconomic variable on the adjustment of the inflation target. u^* and Y_g^* are the target values of unemployment and GDP growth. The adjustment is capped within a reasonable range to prevent extreme inflation rates, specifically between 0% and 8%. This method dynamically adjusts the inflation target based on real-time economic data, such as inflation and unemployment, as well as key economic goals.

AI-Economist AI-Economist [28] is a learning based method using multi-agent reinforcement learning (RL). In this model, agents are driven to maximize their utility, which is a function of savings, consumption, and labor effort. The agents' utility function incorporates consumption and goods prices, while labor exerts a negative influence, reflecting the cost of effort. Specifically, the utility function is given by:

$$U_i = \frac{(s_i/P)^{1-\lambda_s} - 1}{1 - \lambda_s} \times \frac{(c_i/P)^{1-\lambda_c} - 1}{1 - \lambda_c} - \lambda_l l_i \tag{28}$$

where s_i is the agent's savings, c_i is consumption, P is the price of goods, and l_i represents labor effort. The parameters λ_s, λ_c, and λ_l control the relative importance of savings, consumption, and labor, respectively, in determining the agent's utility. The agents in AI-Economist make decisions based on various economic factors, such as monthly wages, interest rates, and tax rates, all of which affect their work and consumption behavior.

The planner in the AI-Economist aims to set tax policies that optimize social welfare, which balances economic productivity and equality. The social welfare function can be expressed as the product of equality and productivity:

$$swf(x^c) = eq(x^c) \cdot prod(x^c) \tag{29}$$

where $eq(x_t)$ represents equality and $prod(x_t)$ refers to the total value of the society, equals to $\sum_{i=1}^{N} x_i^c$. The Planner uses RL to iteratively adjust tax rates over time, based on the agents' evolving behaviors and incomes. It observes macroeconomic indicators and updates the tax schedule accordingly to achieve the desired balance between productivity and equality.

In application, we apply Proximal Policy Optimization (PPO) [20], a reinforcement learning algorithm, to train agents agent planner in maximizing their own utility over time.

A.9 Implementation Parameters

The parameters for the simulation environment, rule-based baselines and AI-Economist are shown below. The total training step for AI-Economist is 24000000.

Experiment Hyperparameters	Values
α_w	0.05
α_p	0.10
r^n	0.01
u^n	0.04
α_π	0.5
α_u	0.5
Y_g^n	0.02
β_π	0.3
β_u	0.2
β_y	0.2
β_{LEN}	0.1
γ_{LEN}	0.1
h_{CATS}	1

Table 1. Additional implementation parameters of simulation environments and rule-based baselines

Training Parameter	**Agent Policy**	**Planner Policy**
clip_param	0.3	0.3
entropy_coeff	0.2	0.025
entropy_coeff_schedule	0: 0.3 10,000,000: 0.1	null
gamma	1.0	0.998
grad_clip	10.0	10.0
kl_coeff	0.0	0.0
kl_target	0.01	0.01
lambda	1.0	0.98
lr	0.0001	0.0003
lr_schedule	null	null
fc_dim	128	256
emb_dim	4	4
emb_vocab	100	100
lstm_cell_size	128	256
num_conv	2	2
num_fc	2	2
max_seq_len	25	25
use_gae	true	true
vf_clip_param	50.0	50.0
vf_loss_coeff	0.05	0.05

Table 2. Training Parameters for AI-Economist

5 Ethics, Fairness, and Normative Reasoning

5.1 Combining Normative Ethics Principles to Learn Prosocial Behaviour

Combining Normative Ethics Principles to Learn Prosocial Behaviour

Jessica Woodgate[0000−0001−9039−846X] and Nirav Ajmeri[0000−0003−3627−097X]

University of Bristol {jessica.woodgate,nirav.ajmeri}@bristol.ac.uk
https://uob-mas.github.io/

Abstract. Prioritising citizens in the design and development of multi-agent systems (MAS) is key to ensuring that MAS are socially beneficial. To ensure that artificial agents within MAS act in citizen-centric ways, agents should foster prosociality, defined as behaving in ways that support the well-being of others. Principles from normative ethics—*the philosophical study of morality*—can be operationalised in the decision-making capacities of agents to discern ethically acceptable actions and promote prosocial behaviour. However, challenges exist in operationalising principles: (1) individual principles may be unintuitive; (2) while incorporating multiple principles mitigates issues with individual principles, conflicts may arise between them. We present PriENE, a method for combining multiple principles in individual decision-making to encourage agents learning prosocial behaviour. We evaluate PriENE in a simulated berry harvesting scenario. Interestingly, preliminary results show that societies of PriENE agents do better along metrics one might expect individual principles to have an advantage: one might expect egalitarianism to minimise inequality, but PriENE societies minimise inequality further; one might expect maximin to have highest minimum experience, but PriENE societies raise minimum experience further; one might expect utilitarianism to have highest cumulative experience, but cumulative experience is further increased in PriENE societies.

Keywords: ethical decision-making · cooperation · fairness · prosociality

1 Introduction

Principles from normative ethics, the rational and systematic study of right and wrong, provide frameworks for guiding moral judgements (Murukannaiah and Singh, 2020; Woodgate and Ajmeri, 2022). Operationalising principles in decision-making enables artificial agents (hereafter referred to as agents) to consider the well-being of others and discern ethically acceptable actions (Woodgate et al., 2025). The capacity of agents to consider others and act ethically becomes paramount as multi-agent systems (MAS) are increasingly adopted in real-world applications with diverse impacts on citizens (Stein and Yazdanpanah, 2023). To ensure that MAS are beneficial to citizens, agents within MAS should be prosocial, defined as acting in ways intended to benefit others (Paiva et al., 2018;

Sundaresan et al., 2025). Implementing principles in decision-making capacities supports agents considering others and learning behaviours that are prosocial insofar as they support the well-being of others as well as the agent's own needs (Ajmeri et al., 2020; Mashayekhi et al., 2022).

Previous works cultivate cooperation and prosociality by appeal to existing behaviours (Anavankot et al., 2023; Dell'Anna et al., 2020; Lupu and Precup, 2020; Tzeng et al., 2024; Santos et al., 2018). However, learning from others without evaluating behaviour to identify potentially better options risks perpetuating existing injustices. Implementing normative ethics mitigates difficulties, as principles are prescriptive, denoting *what ought to happen*, rather than descriptive, denoting *what is happening* (Kim et al., 2021). However, challenges arise with operationalising principles.

(1) Individual principles may be unintuitive. Ethics can be defined in various ways, each with strengths and weaknesses (Woodgate and Ajmeri, 2024). Applying specific principles in certain situations may yield unintuitive outcomes; for instance, utilitarianism, which aims to maximise total utility (Mill, 1863), can lead to unfair treatment of minorities. Using multiple principles in decision-making supports the ability to view problems from diverse perspectives and helps mitigate issues with individual principles.

(2) Principles may conflict. Considering multiple principles broadens ethical reasoning; however, specific principles may conflict with one another. For instance, maximin prioritises improving the minimum experience in society (Rawls, 1967), while egoism seeks the best outcome for oneself (Sidgwick, 1907). Aggregating various principles can help resolve conflicts and balance the strengths and weaknesses of individual recommendations.

Contribution. We present PriENE, a method to operationalise and combine normative ethics principles egoism, utilitarianism, maximin, and egalitarianism in the decision-making of individual agents to learn prosocial behaviours.

Novelty. PriENE advances prior work by (1) implementing a variety of principles in learning mechanisms; (2) aggregating multiple principles to mitigate weaknesses with individual principles.

We empirically evaluate PriENE in a simulated berry harvesting scenario to examine the effects of decision-making in a society with unequal resource distribution. We compare PriENE societies with societies of agents implementing individual principles. Interestingly, we find that PriENE societies do better where one might expect individual principles to have an advantage: PriENE minimises inequality more than egalitarianism; raises minimum experience above maximin; improves total social welfare above utilitarianism.

2 PriENE

We now present the PriENE method. We model PriENE agents using reinforcement learning (RL), in which an agent optimises long-term return by repeatedly interacting with its environment (Sutton and Barto, 2018). A PriENE agent operationalises egoism, which promotes achieving the greatest outcome possible for oneself (Sidgwick, 1907), through basic Q-learning with DQN. DQN is an RL algorithm that uses a neural network to parametrise an approximate Q-function (Mnih et al., 2015).

To consider well-being of others and learn prosocial behaviour that is citizen beneficial, a PriENE agent operationalises normative ethics. We adapt the utility function proposed by Leben (2020) to model a distribution of resources d and well-being of each member of society. From Leben (2020), $u_i(d) \rightarrow (\upsilon_i)$ models a distribution of resources d for an agent i; n is the number of living agents; (υ_i) is a measurement of well-being for each agent $ag_1, \ldots, ag_n$; $u_t(d, \upsilon_i)$ is utility for agent i given its resources d at time t; $U_t = \{u_t(d, \upsilon_1), \ldots, u_t(d, \upsilon_n)\}$ is the set of utilities for all agents in a society at t. To operationalise each principle, a PriENE agent compares U_t, before acting and U_{t+1}, after acting. A sanction is a reaction to approved or disapproved behaviour. A PriENE agent perceives a self-directed sanction f (directed towards and affecting only its sender (Nardin et al., 2016)) from each principle $p_1, \ldots, p_m$ indicating whether utility improved, worsened, or did not change.

Utilitarianism. Maximise total net utility (Mill, 1863). Compute utility distributions by summing aggregate utilities, thus $UT = \sum_{i=1}^{n} u(d, \upsilon_i)$.

Maximin. Prioritise well-being of the worst-off (Rawls, 1967). Compute minimum experience–lowest utility of an agent, $MA = min_i u(d, \upsilon_i)$.

Egalitarianism. Confer equal shares to each individual (Binns, 2018). Compute accumulated difference of each agent's utility to an ideal where all agents are perfectly equal. Thus, $EG = \sum_{i=1}^{n} |u(d, \upsilon_i) - \mu(U)|$ where $\mu(U) = \frac{\sum_{i=1}^{n} u(d,\upsilon_i)}{n}$ denotes average utility of the society.

Aggregating principles mitigates difficulties with individual principles. A PriENE agent computes aggregated sanction F from mean of all sanctions $f_{p_1}, \ldots, f_{p_m}$ so that $F(f_{p_1}, \ldots, f_{p_m}) = \frac{1}{m} \sum_{i=1}^{m} f_{p_i}$. Various ways of combining principles may be appropriate for distinct scenarios. For example, in a situation where causing harm is risky, if even one principle indicates an action is negative, the overall sanction should be negative. In scenarios where it is preferable to incorporate the recommendations of all principles, the overall sanction could be computed as the mean of all sanctions.

To make decisions, at each time step t, a PriENE agent observes state s_t and selects action a with predicted max Q-value from DQN. After acting, the agent perceives reward r from the environment. For each principle, the agent calculates self-directed sanction $f_{p_1}, \ldots, f_{p_m}$, and aggregates them to obtain sanction F. The agent combines F with the environment reward r through reward shaping, which provides immediate feedback based on heuristics, resulting in $r' = r + F$. Finally, the agent passes r' to DQN for learning.

3 Experimental Setup

We create a harvest environment, illustrated in Figure 1, in which an agent can move, forage for berries, eat berries, throw berries to other agents. To examine the effects of various principles, we train five agent types: egoistic, egalitarian, maximin, utilitarian, and PriENE. We run $e = 1000$ episodes. Each episode runs until all agents have died or $t_{\max} = 200$ steps. Appendix A.1 includes additional details of experimental setup including compute information, hyperparameter selection, simulation parameters, and range of values tried.

Reproducibility Our simulation codebase, including complete simulation parameters, is publicly available (Woodgate and Ajmeri, 2025).

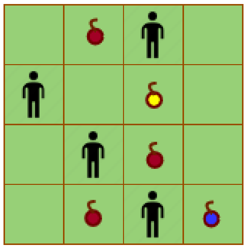


Fig. 1. Colours harvest. Each agent moves freely but only collects berries of a specific colour. Some colours are more plentiful, giving those agents access to more resources. Agents can throw berries to each other across the grid.

Metrics We examine the quality of individual agents' experience, measured by the number of berries consumed ag_{berries}. To evaluate fairness, we compute:
$\mathbf{M_1}$ (inequality). Gini index (distance to perfect equality (Gini, 1912)) of accumulated ag_{berries} across the society. Lower is better.
$\mathbf{M_2}$ (minimum experience). Minimum individual accumulated ag_{berries} at the end of each episode. Higher is better.
To evaluate sustainability, we assess the following metrics:
$\mathbf{M_3}$ (maximum experience). Maximum individual accumulated ag_{berries} at the end of each episode. Higher is better.
$\mathbf{M_4}$ (social welfare). Total ag_{berries} accumulated at the end of each episode. Higher is better.
$\mathbf{M_5}$ (robustness). Length of each episode. Higher is better.

4 Results

Table 1 displays preliminary results of ag_{berries} mean for PriENE societies and societies implementing individual principles. Results are calculated by examining the ag_{berries} of each individual agent at the end of each episode.

M_1 (inequality) is lowest in PriENE societies and highest in utilitarian societies. M_2 (minimum experience) is highest egalitarian followed by PriENE. M_3 (maximum experience) is highest in utilitarian societies, followed by maximin, egoistic, PriENE, then egalitarian. M_4 (social welfare) is highest in maximin societies followed by PriENE. M_5 (robustness) is highest in PriENE societies.

Table 1. Comparing PriENE and individual principles mean $\bar{x}$ and standard deviation σ of ag_{berries}.

Metric	Egoistic		Utilitarian		Maximin		Egalitarian		PriENE	
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$	σ
M_1	0.43	0.14	0.48	0.11	0.42	0.12	0.38	0.14	0.37	0.11
M_2	2.06	2.43	1.96	2.15	2.09	2.14	3.23	4.69	2.81	2.59
M_3	35.48	11.66	42.9	5.72	37.95	9.5	29.28	12.05	35.0	11.47
M_4	69.12	34.08	77.14	21.9	79.51	29.06	65.82	39.13	78.64	34.19
M_5	95.08	83.91	100.55	83.75	106.2	85.38	94.83	79.48	106.85	82.12

Discussion Interesting highlights include: one might expect egalitarian to minimise inequality but PriENE minimises inequality further; one might expect maximin to have highest minimum experience but PriENE improves minimum more; one might expect utilitarian to have highest social welfare but PriENE is higher. The combination of lowest inequality and second highest minimum experience indicates PriENE societies have a satisfactory level of fairness. PriENE societies also have second highest social welfare and highest robustness, indicating a PriENE agent is not acting at unlimited cost to itself; it is not giving away berries when it is itself in need. Results suggest PriENE agents learn prosocial behaviours that support the well-being of the agent as well as the society.

5 Conclusions and Directions

PriENE is a method for operationalising multiple normative ethics principles in decision-making capacities of individual agents, to promote prosocial and citizen-centric behaviour. Overall, results show that PriENE societies lead to lowest inequality, second highest minimum experience and social welfare, and highest robustness. These results suggest that PriENE encourages agents to learn prosocial behaviours that support the well-being of others, fostering the propensity of MAS to benefit the interests of citizens. To expand analysis to more complex settings, future directions involve evaluating heterogeneous societies in which agents operationalise different principles to one another; implementing scenarios closer to the real world; increasing the agent population; inferring well-being of others utilising solely local information, such as identifying implicit responsibility (Chopra et al., 2024); examining the effects of limited observability; considering longer term impact of actions on the well-being of others; exploring the influence of context on ethical decision-making including social norms, which are standards of expected behaviour (Chopra et al., 2018; Morris-Martin et al., 2019); implementing additional principles (Woodgate and Ajmeri, 2024).

Acknowledgements This work was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/W524414/1, EP/Y028392/1].

Bibliography

Nirav Ajmeri, Hui Guo, Pradeep K. Murukannaiah, and Munindar P. Singh. 2020. Elessar: Ethics in Norm-Aware Agents. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. IFAAMAS, Auckland, 16–24.

Amritha Menon Anavankot, Stephen Cranefield, and Bastin Tony Roy Savarimuthu. 2023. Towards Norm Entrepreneurship in Agent Societies. In *Advances in Practical Applications of Agents, Multi-Agent Systems, and Cognitive Mimetics. The PAAMS Collection*. Springer, Switzerland, 188–199.

Yoshua Bengio. 2012. Practical Recommendations for Gradient-Based Training of Deep Architectures. In *Neural Networks: Tricks of the Trade: Second Edition*. Springer, Berlin, 437–478

Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proc. FAccT*, Vol. 81.

Amit Chopra, Torre Leendert Van Der, and Harko Verhagen. 2018. *Handbook of Normative Multiagent Systems*. College Publications, Rickmansworth, UK.

Daniel Collins, Conor Houghton, and Nirav Ajmeri. 2024. *Proceedings of the 2nd International Workshop on Citizen-Centric Multi-Agent Systems (CMAS)*. College Publications, Auckland.

Davide Dell'Anna, Mehdi Dastani, and Fabiano Dalpiaz. 2020. Runtime Revision of Sanctions in Normative Multi-Agent Systems. *Autonomous Agents and Multi-Agent Systems (JAAMAS)* 34, 2 (2020), 1–54.

Corrado Gini. 1912. *Variabilità e mutabilità : contributo allo studio delle distribuzioni e delle relazioni statistiche*. Universitá di Cagliari, Cagliari.

Tae Wan Kim, John Hooker, and Thomas Donaldson. 2021. Taking Principles Seriously: A Hybrid Approach to Value Alignment in Artificial Intelligence. *JAIR* 70 871–890.

Derek Leben. 2020. Normative Principles for Evaluating Fairness in Machine Learning. In *Proc. AIES*. ACM, New York, 86–92.

Andrei Lupu and Doina Precup. 2020. Gifting in Multi-Agent Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. IFAAMAS, Auckland, 789–797.

Mehdi Mashayekhi, Nirav Ajmeri, George F. List, and Munindar P. Singh. 2022. Prosocial Norm Emergence in Multiagent Systems. *ACM TAAS* 17, 3:1–3:24.

John S. Mill. 1863. *Utilitarianism*. Longmans, Green and Company.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540, 529–533.

Andreasa Morris-Martin, Marina De Vos, and Julian Padget. 2019. Norm Emergence in Multiagent Systems: A Viewpoint Paper. *JAAMAS* 33, 6, 706–749.

Pradeep K. Murukannaiah and Munindar P. Singh. 2020. From Machine Ethics to Internet Ethics: Broadening the Horizon. *IEEE IC* 24, 3, 51–57.

Luis G. Nardin, Tina Balke-Visser, Nirav Ajmeri, Anup K. Kalia, Jaime S. Sichman, and Munindar P. Singh. 2016. Classifying Sanctions and Designing a Conceptual Sanctioning Process Model for Socio-Technical Systems. *KER* 31, 142–166.

Ana Paiva, Fernando Santos, and Francisco Santos. 2018. Engineering Pro-Sociality With Autonomous Agents. *Proc. AAAI* 32, 1, 7994–7999.

John Rawls. 1967. Distributive Justice. *Philosophy, Politics and Society* 1, 58–82.

Fernando P. Santos, Jorge M. Pacheco, and Francisco C. Santos. 2018. Social norms of cooperation with costly reputation building. In *Proc. AAAI.* AAAI Press, New Orleans, Article 579, 8 pages.

Henry Sidgwick. 1907. *The Methods of Ethics.* Macmillan Publishers, London.

Sebastian Stein and Vahid Yazdanpanah. 2023. Citizen-Centric Multiagent Systems. In *Proc. AAMAS.* IFAAMAS, London, 1802–1807 pages.

Divya Sundaresan, Akhira Watson, Eleni Bardaka, Crystal Chen Lee, Christopher B. Mayhorn, and Munindar P. Singh. 2025. Prosociality in Microtransit. *JAIR* 82 (2025), 34.

Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement learning: An introduction* (second edition ed.). The MIT Press, Cambridge, Massachusetts.

Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. 2024. Norm Enforcement with a Soft Touch: Faster Emergence, Happier Agents. In *Proc. AAMAS.* IFAAMAS, Auckland, 1837–1846.

Jessica Woodgate and Nirav Ajmeri. 2022. Macro Ethics for Governing Equitable Sociotechnical Systems. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS).* IFAAMAS, Online, 1824–1828. Blue Sky Ideas Track.

Jessica Woodgate and Nirav Ajmeri. 2024. Macro Ethics Principles for Responsible AI Systems: Taxonomy and Directions. *CSUR* 56, 289 (July 2024), 1–37.

Jessica Woodgate and Nirav Ajmeri. 2025. Codebase for Combining Normative Ethics Principles to Learn Prosocial Behaviour. `https://doi.org/10.5281/zenodo.14884503`.

Jessica Woodgate, Paul Marshall, and Nirav Ajmeri. 2025. Operationalising Rawlsian Ethics for Fairness in Norm-Learning Agents. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI).* AAAI, Philadelphia, 1–9.

A Appendix

A.1 Details of Experimental Setup

We train $k = 4$ agents in a grid size 8×8 with $b_{\text{initial}} = 12$ berries. An agent begins with $h_{\text{initial}} = 5.0$ health, which decays by $h_{\text{decay}} = 0.1$ at each time step. If an agent forages at a location where a berry exists then it can carry the berry in its bag $ag_{\text{berries-carried}}$. If an agent has $b >= 1$ in its bag, it can eat a berry and its health increases by h_{gain}. When an agent's health is above h_{throw}, it can throw a berry to another agent. For simplification, agents observe one another's well-being as a single number $ag_{\text{well-being}}$. $ag_{\text{well-being}}$ is measured by how many days an agent has left to live, a function of how many berries it is carrying and its health, where $ag_{\text{well-being}} = \frac{ag_{\text{health}}+(ag_{\text{berries-carried}} \times h_{\text{gain}})}{h_{\text{decay}}}$. In the real world, this is analogous to agents observing a sickly appearance. Each episode runs until all agents die, or $t_{\max} = 200$ steps.

A.2 Computing Infrastructure

We conducted the simulation experiments on a workstation with Intel Xeon Processor W-2245 (8C 3.9 GHz), 256GB RAM, and Nvidia RTA A6000 48GB GPU.

A.3 Hyperparameter Selection

Table 2 lists the interaction module parameters and range of values tried per parameter. We select these parameters empirically, with reference to literature Bengio (2012).

Table 2. DQN Parameters.

Description	Parameter	Range Tried	Final Value	Criterion
Batch size	B	{32, 64, 128}	64	Training time
Iteration for updating weights of target network	C	{1000, 100, 50}	50	Test performance
Probability of exploration	ϵ	0.9–0.0	0.0	Test performance
Learning rate	α	{0.01, 0.001, 0.0001}	0.0001	Test performance
Number of hidden units	Hn	{32, 64, 128}	128	Test performance
Number of hidden layers	Hl	1–3	2	Test performance

A.4 Simulation Parameters Selection

Table 3 lists the simulation parameters and range of values tried per parameter. We select these parameters empirically.

Table 3. Simulation Parameters.

Description	Parameter	Range Tried	Final Value
Grid size	$o_{\text{basic}} \times p_{\text{basic}}$	$\{4 \times 4,\ 8 \times 4\}$	8×8
Number of agents	k	{2, 4}	4.0
Initial number of berries	b_{initial}	{8, 12, 16}	12.0
Initial health of agent	h_{initial}	{5.0, 10.0}	5.0
Health decay	h_{decay}	{–0.01, –0.1}	−0.01
Health gain from eating berry	h_{gain}	{0.1, 1.0}	0.1
Minimum health to throw	h_{throw}	{0.5, 0.6, 1.0}	0.6
Number of episodes	e	{500, 1000}	1000.0
Maximum steps in episode	$t_{\max}$	{50, 200}	200.0

Rewards To encourage agents to learn to survive, agents are positively rewarded for reaching the end of an episode and negatively rewarded for dying. Agents are rewarded for throwing to provide incentive for egoistic agents that do not implement ethics sanctioning to learn cooperative behaviours. Providing environmental rewards for cooperative behaviours allows for fair comparison between agent types, rather than giving agents that implement ethical principles additional rewards. Rewards are also normalised between egoistic agents and agents implementing ethical principles to avoid obvious results by giving additional rewards. Table 4 lists complete rewards received by agents.

Table 4. Rewards received by an agent. Rewards are normalised between egoistic and other agents to avoid obvious results by giving agents implementing other principles more rewards.

Action	Egoistic	Principles
Survive episode	1.00	1.00
Eat berry	1.00	0.80
Forage berry	1.00	0.80
Throw berry	0.50	0.50
Try to eat without berries	−0.20	−0.10
Try to throw without berries	−0.20	−0.10
Try to throw without sufficient health	−0.20	−0.10
Try to throw without recipient	−0.20	−0.10
Die	−1.00	−1.00
Positive ethics sanction	0.00	0.40
Negative ethics sanction	0.00	−0.40

5.2 Quantifying the Self-Interest Level of Markov Social Dilemmas

Quantifying the Self-Interest Level of Markov Social Dilemmas

Richard Willis[1][0009−0007−7909−8491], Yali Du[1][0000−0001−5683−2621], Joel Z Leibo[1,2][0000−0002−3153−916X], and Michael Luck[3][0000−0002−0926−2061]

[1] King's College London, London, UK
richard.willis@kcl.ac.uk
[2] Google DeepMind, London, UK
[3] University of Sussex, Brighton, UK

Abstract. This paper introduces a novel method for estimating the self-interest level of Markov social dilemmas. We extend the concept of self-interest level from normal-form games to Markov games, providing a quantitative measure of the minimum reward exchange required to align individual and collective interests. We demonstrate our method on Commons Harvest, which represents a common-pool resource.

Keywords: Social Dilemma · Game Theory · Reinforcement Learning.

1 Introduction

Social dilemmas are situations where individual incentives conflict with group interests, and they present significant challenges in multiagent cooperation. Agents need sufficient motivation to care about others for collective action to become more attractive than selfish behaviour. We address this with reward exchange, whereby agents agree to exchange a fixed proportion of their rewards with each other, creating an incentive for them to improve the well-being of others.

The self-interest level [11] quantifies the greatest proportion of their own rewards that agents can retain while using reward exchange to resolve a social dilemma. It serves as a solution to social dilemmas, and a metric for players' propensity to cooperate, assessing the gap between individual and collective incentives. A low self-interest level indicates strong incentives for players to avoid prosocial behaviour. In this paper we present a novel method for estimating the self-interest level of stochastic game representations of social dilemmas.

2 Related Work

Our work focuses on extending two prominent approaches. The first approach develops game-theoretic metrics to quantify the amount of shared interest required to achieve socially optimal equilibria in mixed-motive games [1,5,3,2]. These contributions are applicable to analytically tractable games. The second approach [9,7] develops more complex models using stochastic games that can capture nuanced aspects of real-world social dilemmas, such as cooperativeness as a graded quantity, agents with partial information about the state of the world, and decisions with temporally extended consequences. In this setting, reward transfers have been used to promote collective behaviours in multiagent reinforcement learning agents [10,6,12].

3 Background

3.1 Markov Social Dilemmas

Social dilemmas are situations in which individuals face the choice between acting selfishly (to defect) for personal gain or acting in a prosocial manner (to cooperate) which yields greater overall benefits to the collective. For all agents: (i) the collective does better when an agent chooses to cooperate than when the agent chooses to defect; (ii) each agent may be better off individually when it defects; and, (iii) all agents prefer mutual cooperation over mutual defection.

An n-player Markov social dilemma is a tuple $(M, \overrightarrow{\Pi} = \overrightarrow{\Pi}_c \cup \overrightarrow{\Pi}_d)$, where M is a Markov game and $\overrightarrow{\Pi}_c$ and $\overrightarrow{\Pi}_d$ are two disjoint sets of policies said to implement cooperation and defection respectively. The utility function for player i, $R_i(\overrightarrow{\pi})$, denotes the expected total reward in a game rollout given the joint policies, and satisfies the following properties:

$$\begin{aligned}
&\text{(i)} \quad \forall i \quad \textstyle\sum_j R_j(\pi_c{}^\frown\overrightarrow{\pi_{-i}}) > \sum_j R_j(\pi_d{}^\frown\overrightarrow{\pi_{-i}}) \\
&\text{(ii)} \quad \forall i \quad \exists \overrightarrow{\pi_{-i}} : \; R_i(\pi_d{}^\frown\overrightarrow{\pi_{-i}}) > R_i(\pi_c{}^\frown\overrightarrow{\pi_{-i}}) \\
&\text{(iii)} \quad \forall i \quad R_i((\pi_c, \pi_c \dots \pi_c)) > R_i((\pi_d, \pi_d \dots \pi_d))
\end{aligned}$$

Where $\overrightarrow{\pi_{-i}}$ represents the tuple of policies for all players other than player i, and $^\frown$ is a coupling operator that inserts π_i into $\overrightarrow{\pi_{-i}}$ such that $\overrightarrow{\pi} = \pi_i{}^\frown\overrightarrow{\pi_{-i}}$.

3.2 Reward Exchange

We allow the agents to enter into a contract to exchange proportions of their future rewards between one another. We introduce a parameter, s, denoting the proportion of its own rewards that an agent retains, termed the *self-interest* of the agents. The remainder, $1 - s$, is distributed equally among the other $n - 1$ co-players. The post-transfer reward function for agent i, R'_i, is therefore:

$$R'_i(\overrightarrow{r}, s) = sr_i + \frac{1-s}{n-1} \sum_{j \neq i} r_j \tag{1}$$

3.3 Self-Interest Level

We say that a social dilemma is resolved when all agents prefer to cooperate. The self-interest level of a Markov social dilemma, denoted s^*, is defined as:

$$s^* = \max\{s \mid \forall i \; R'_i(R(\pi_C{}^\frown\overrightarrow{\pi_{-i}}), s) > R'_i(R(\pi_D{}^\frown\overrightarrow{\pi_{-i}}), s)\}$$

Note that when $s = \frac{1}{n}$, the post-transfer reward function (Equation (1)) for all agents is equivalent to maximising the sum of rewards:

$$R'_i(\overrightarrow{r}, \frac{1}{n}) = \frac{1}{n}\sum_{j=1}^{n} r_j$$

Consequently, in a Markov social dilemma with $s = \frac{1}{n}$, cooperative policies are preferable to defect policies, because cooperation increases the total reward due to inequality (i). However, players may still strictly prefer cooperative policies for $s > \frac{1}{n}$. The self-interest level therefore has a lower bound, and $s^* \in [\frac{1}{n}, 1]$.

4 Method

We now introduce a method to estimate the self-interest level of Markov social dilemmas. Computing dominant policies is computationally intractable, so we use learning algorithms to find approximately optimal joint policies, and assesses the resulting equilibria.

4.1 Estimating the self-interest level

We consider joint policies achieving equivalent collective reward to those trained to maximise the sum of rewards (when $s = \frac{1}{n}$) to be maximally cooperating. The self-interest level is estimated as the largest value of s for which independent policies converge to a maximally cooperative equilibrium.

Writing joint policies trained with a self-interest of s as $\overrightarrow{\pi_s}$:

$$s^* = \max_{\frac{1}{n} \leq s \leq 1} : \sum\nolimits_j R_j(\overrightarrow{\pi_s})) \approx \sum\nolimits_j R_j(\overrightarrow{\pi_{\frac{1}{n}}}))$$

We wish to guarantee that policies convergence to cooperative equilibria, regardless of their initialisation, implying that cooperation is dominant. We approximate this by choosing challenging initialisations: policies that have converged to equilibria with poor rewards, found by training without reward exchange, so that the agents have incentives to act selfishly and shirk cooperation.

4.2 Policy Training

The policies are trained in two stages:

1. **Pre-training:** We gradually increase the number of players in the environment, training for a fixed number of episodes each time.
2. **Training:** We continue training the independent policies while iteratively decreasing their self-interest after a number of episodes.

We use a range of self-interest values based on the ratio of the fraction of reward an agent keeps for itself compared to the proportion of a co-players' reward it receives, because the agents typically face a choice between taking a benefit for themselves, or allowing a co-player to gain it. The ratios we use are [20:1, 10:1, 5:1, 3:1, 5:2, 2:1, 5:3, 4:3, 1:1].

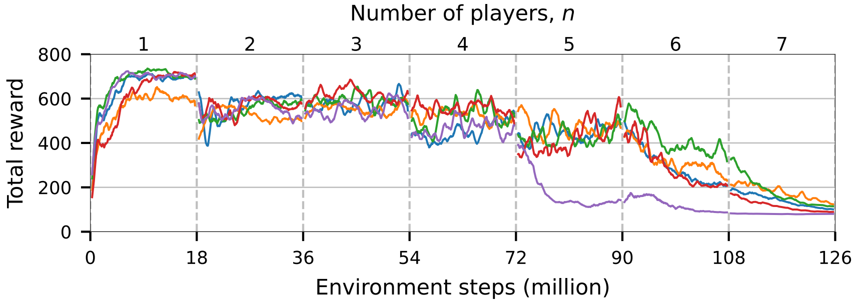


Fig. 1: Pre-training, increasing numbers of players

4.3 Evaluation

Due to the stochastic nature of reinforcement learning and Markov games, we repeat the training for five different policy initialisations. Furthermore, we identify the self-interest level with a degree of tolerance, selecting the largest value of s that achieves a total reward not statistically worse than the best measured:

- Compute the mean and standard deviation of the sum of rewards after training for each value of s, and set $s_{\max}$ to the s value with the largest mean.
- Conduct a one-sided Dunnett's test [4], a method to compare multiple samples with a single control, to assess which of the means are statistically worse than $s_{\max}$. We accept a p-value of < 0.1 as significant.
- Choose s^* as the largest s value with a total reward mean that is not statistically worse that that of $s_{\max}$, otherwise $s^* = s_{\max}$ if all are worse.

5 Results

We illustrate this process on Commons Harvest [8], which models a common pool resource. The challenge is to manage the resource sustainably and avoid a tragedy of the commons. Commons Harvest comprises seven agents harvesting apples from four large and two small apple patches. Collecting an apple provides as reward of 1. Harvested apples regrow with probability proportional to the number of apples remaining in the patch. If all the apples in a patch are harvested, however, it is depleted and no apples will regrow.

5.1 Pre-training

Figure 1 shows that the best performance is achieved when there is only a single agent. In principle, multiple players should be able to match or exceed the reward of a single agent. That they fail to do so in practice is due to the mixed-motive structure of the rewards for $n > 1$ players. While the benefits of harvesting an apple are entirely captured by the harvester, the cost of a reduced regrowth rate is shared among all agents; consequently, the agents have incentives to harvest

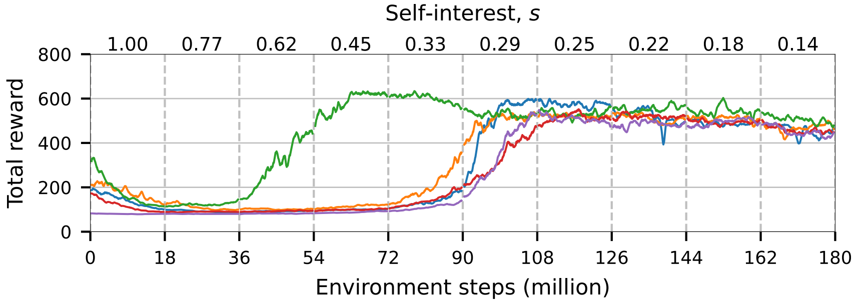


Fig. 2: Iteratively decreasing self-interest during training

s	1	0.77	0.63	0.46	0.33	**0.29**	0.25	0.22	0.18	0.14
mean	102	102	169	208	295	**540**	510	524	519	464
std dev	20	28	165	231	171	45	48	30	13	24
p-value	0	0	0	0.01	0.01	N/A	0.17	0.27	0.17	0.01

Table 1: Dunnett's test results

more apples than is socially optimal. Over the range of 2–4 players, all five seeds maintain good social outcomes. However, with 5–7 players, the performance collapses to poor equilibria. Here, the agents quickly consume every apple, so nothing is able to regrow, and the tragedy of the commons has materialised.

5.2 Training

Figure 2 demonstrates that all seeds have recovered good social performance by $s = 0.29$. Although the agents achieve a reward slightly less than that which a single agent can achieve (in Figure 1, with $n = 1$), this is due to the difficulty of coordination in independent multiagent reinforcement learning. Indeed, by $s = 0.14$, the agents have an effective team reward, and would be better off if one of them followed the single agent policy and the other six remained inactive.

We calculate the mean and standard deviation of the total reward achieved at the end of training for all values of s, and compute the one-sided Dunnett's test p-value to determine whether the means are statistically lower. The results are presented in Table 1. In this case, the optimal performing value is $s = 0.29$, and all the larger values of s have a statistically worse mean total reward. We therefore estimate the self-interest level to be in the range $0.29 \leq s^* \leq 0.33$. Longer training periods or a larger number of seeds might yield slightly different results, by limited computational resources prevented exhaustive investigation. To validate the self-interest level for Commons Harvest, we train new policies with fixed reward exchange proportions: $s = 1$ (fully independent), $s = s^*$ (self-interest level), $s = \frac{1}{n}$ (team reward) and a value of s slightly larger than the range that s^* was determined to lie within, which we call s^+.

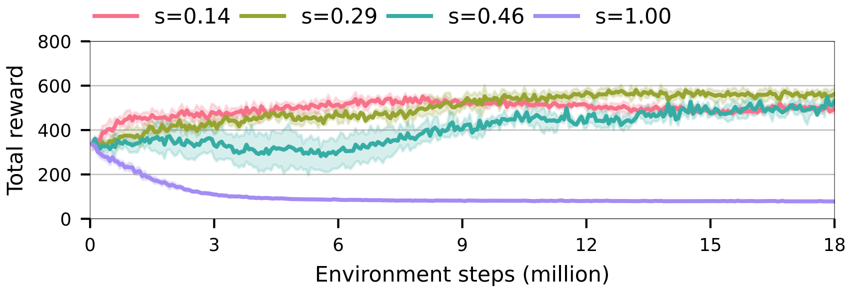


Fig. 3: Training with fixed self-interest

Figure 3 confirms that without reward exchange, policies fail to learn cooperation, while training at the self-interest level reaches a cooperative equilibrium that slightly outperforms the team reward and s^+. This demonstrates how our method enables cooperation without requiring agents to completely sacrifice their individual interests. By using the minimal necessary reward exchange rather than full team rewards, we potentially avoid credit assignment challenges while still achieving socially optimal outcomes.

6 Conclusion

We introduced a novel method for estimating the self-interest level of Markov social dilemmas, bridging the gap between game-theoretic metrics and complex multiagent reinforcement learning models. The self-interest level serves as a valuable metric for assessing the propensity of cooperation in mixed-motive games, by quantifying the gap between individual and collective incentives.

Our work offers both practical metrics and solutions for real-world social dilemmas. As a metric, the self-interest level enables risk assessment: systems with low self-interest levels face significant barriers to cooperation and may be prone to conflict, allowing system designers to identify where intervention is necessary. As a solution mechanism, reward exchange can be applied to problems like fishery management, where traditional quotas often fail due to persistent incentives to overfish. If fishing nations were to exchange a proportion of their fishing profits with other fishing nations, this would reduce each country's incentive to overexploit fish stocks while simultaneously motivating all participating countries to improve ocean health.

Acknowledgments. This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org) and a BT/EPSRC funded iCASE Studentship [grant number EP/T517380/1].

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Apt, K.R., Schaefer, G.: Selfishness Level of Strategic Games. Journal of Artificial Intelligence Research **49**, 207–240 (Feb 2014). https://doi.org/10.1613/jair.4164
2. Caragiannis, I., Kaklamanis, C., Kanellopoulos, P., Kyropoulou, M., Papaioannou, E.: The Impact of Altruism on the Efficiency of Atomic Congestion Games. In: Wirsing, M., Hofmann, M., Rauschmayer, A. (eds.) Trustworthly Global Computing, vol. 6084, pp. 172–188. Springer Berlin Heidelberg, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15640-3_12
3. Chen, P.A., De Keijzer, B., Kempe, D., Schäfer, G.: The Robust Price of Anarchy of Altruistic Games. In: Chen, N., Elkind, E., Koutsoupias, E. (eds.) Internet and Network Economics, vol. 7090, pp. 383–390. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25510-6_33
4. Dunnett, C.W.: A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association **50**(272), 1096–1121 (1955). https://doi.org/10.2307/2281208
5. Elias, J., Martignon, F., Avrachenkov, K., Neglia, G.: Socially-Aware Network Design Games. In: 2010 Proceedings IEEE INFOCOM. pp. 1–5. IEEE, San Diego, CA, USA (Mar 2010). https://doi.org/10.1109/INFCOM.2010.5462275
6. Gemp, I., McKee, K.R., Everett, R., Duéñez-Guzmán, E.A., Bachrach, Y., Balduzzi, D., Tacchetti, A.: D3C: Reducing the Price of Anarchy in Multi-Agent Learning. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems. pp. 498–506. International Foundation for Autonomous Agents and Multiagent Systems, Online (Feb 2022). https://doi.org/10.5555/3535850.3535907
7. Hughes, E., Leibo, J.Z., Phillips, M., Tuyls, K., Dueñez-Guzman, E., Castañeda, A.G., Dunning, I., Zhu, T., McKee, K., Koster, R., Roff, H., Graepel, T.: Inequity aversion improves cooperation in intertemporal social dilemmas. In: 32nd Conference on Neural Information Processing Systems. pp. 3330–3340. Curran Associates, Inc., Montréal, Canada (2018)
8. Leibo, J.Z., Duéñez-Guzmán, E., Vezhnevets, A.S., Agapiou, J.P., Sunehag, P., Koster, R., Matyas, J., Beattie, C., Mordatch, I., Graepel, T.: Scalable Evaluation of Multi-Agent Reinforcement Learning with Melting Pot. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 6187–6199. PMLR (Jul 2021)
9. Leibo, J.Z., Zambaldi, V., Lanctot, M.: Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In: Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems. pp. 464–473. ACM, São Paulo, Brazil (May 2017)
10. Schmid, K., Kölle, M., Matheis, T.: Learning to Participate through Trading of Reward Shares. In: Proceedings of the 15th International Conference on Agents and Artificial Intelligence. vol. 1, pp. 355–362. SCITEPRESS, Lisbon, Portugal (Feb 2023). https://doi.org/10.5220/0011781600003393
11. Willis, R., Du, Y., Leibo, J.Z., Luck, M.: Resolving social dilemmas with minimal reward transfer. Autonomous Agents and Multi-Agent Systems **38**(2), 49 (Oct 2024). https://doi.org/10.1007/s10458-024-09675-4
12. Yi, Y., Li, G., Wang, Y., Lu, Z.: Learning to Share in Multi-Agent Reinforcement Learning. In: Proceedings of the 36th Conference on Neural Information Processing Systems. Curran Associates Inc. (2022)

A Appendix

See http://arxiv.org/abs/2501.16138 for our full paper, which features greater discussion, further experiments and additional environments. More detail on our experimental setup can be found in our GitHub at https://github.com/willis-richard/meltingpot/tree/markov_sd/.

5.3 Citizen Strategies in School Choice: How Strategic Agents Influence Rank-Minimizing Matchings

Citizen Strategies in School Choice: How Strategic Agents Influence Rank-Minimizing Matchings

Mayesha Tasnim[1][0000−0002−0127−4797], Youri Weesie[2][0009−0001−9895−566X], Sennay Ghebreab[1][0009−0007−5788−4635], and Max Baak[3][0009−0004−0055−5476]

[1] University of Amsterdam
{m.tasnim,s.ghebreab}@uva.nl
[2] Department of Econometrics, Vrije Universiteit Amsterdam
y.c.a.weesie@vu.nl
[3] ING Analytics, ING Bank
m.baak@ing.com

Abstract. We consider one-sided matching problems in citizen-facing allocation systems such as school choice. In these settings, agents are allocated items based on their stated preferences. Posing this as an assignment problem, the average rank of obtained matchings can be minimized using the Rank Minimization (RM) mechanism. RM matchings can lead to significantly better rank distributions than matchings obtained by random-priority mechanisms such as Random Serial Dictatorship (RSD). However, these matchings are also vulnerable to strategic behavior, where agents manipulate their reported preferences to achieve better outcomes. In this work, we derive a best response strategy for a scenario where agents aim to be matched to their top-n preferred items using the RM mechanism under a simplified cost function. This strategy is then extended to a first-order heuristic strategy for being matched to the top-n items in a setup that minimizes the average rank. Based on this finding, an empirical study is conducted examining the impact of the first-order heuristic strategy. The study utilizes data from both simulated markets and real-world matching markets in Amsterdam, taking into account variations in item popularity, fractions of strategic agents, and the preferences for the n most favored items. For most scenarios, RM yields more rank efficient matches than Random Serial Dictatorship, even when agents apply the first-order heuristic strategy. In competitive markets, the matching performance can become worse when 50% of agents or more want to be matched to their top-1 or top-2 preferred items and apply the first-order heuristic strategy to achieve this. These findings contribute to the design of matching systems, showing how agents might manipulate preferences and how this manipulation can impact allocation efficiency.

Keywords: Matching · Strategic Manipulation · Rank Minimization

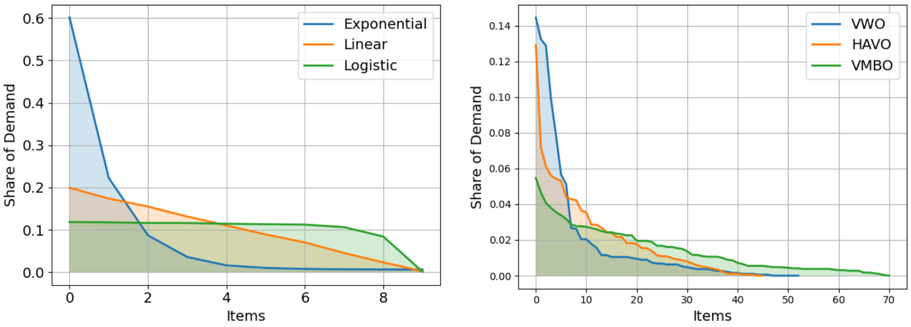
1 Introduction

Matching agents to items given agents' preferences is an essential problem with real-world applications such as school admissions and housing allocation. [6, 7]. Deferred Acceptance with Single Tie-Breaking (DA-STB) is the most well-known matching algorithm, providing stable, envy-free and Pareto-optimal matchings for two-sided preferences [14]. When preferences are one-sided, DA-STB reduces to Random Serial Dictatorship (RSD), and matches are no longer efficient [3, 8]. This inefficiency is reflected in the rank distributions [10] commonly reported by institutions that apply matching mechanisms [1, 2]. Recent works have proposed the rank-minimizing (RM) mechanism [4,10,13,17], which minimizes the average rank received by all agents. Despite efficiency gains, implementing RM in the real world is risky as it is not strategyproof, and agents can receive better matches by misreporting their preferences [4, 17]. In matching problems with one-sided preferences, it is impossible for a mechanism to provide more efficient matches compared to RSD without being vulnerable to manipulation [15]. [17] shows that although RM is manipulable, it is not an *obviously* manipulable mechanism, as no single strategy ensures beneficial gains over being truthful without complete knowledge of all other agents' preferences; suggesting that the shortcoming of non-strategyproofness in RM may not be so severe. [13] show through an empirical study that when agents are strategic using i.i.d. preferences, they do not stand to gain significantly better allocations. However, this study assumes that agents misreport preferences uniformly, and has a uniform distribution of preferences over items.

We motivate that agents can be strategic despite not having complete information of others' preferences, and the impact of strategic preferences can vary across markets with differing demand for items. This extended abstract summarizes our work which implements rank minimization using the well-known Hungarian Algorithm (RM-HAL), assuming linear cost over rank [16]. Our contributions are as follows: we derive a best response strategy for RM-HAL under a simplified cost function, propose a heuristic strategy when costs are linear, and measure the impact of these strategies on matching performance across various market conditions. We find that RM-HAL provides better rank efficiency than RSD, particularly when the number of strategic agents is limited. However, performance declines when more than half of the agents apply the heuristic strategy to secure top-1 or top-2 matches.

2 Strategies for RM-HAL

Let S and M denote sets of agents and items. P represents the set of all possible ordered preference lists, where each list has a fixed length l and $p_{i,n}$ denotes the n^{th} preferred item of agent i ($n \in \mathbb{Z}^+$). Let n-rank popularity of an item $f_{j,n}$ be denoted by the difference between the total number of agents picking item j within their top-n ranked choices and the capacity of j. Items with $f_{j,n} > 0$ are considered popular, those with $f_{j,n} \leq 0$ are less popular. Let N be the set of all popular items.



(a) Synthetic data.

(b) Amsterdam data.

Fig. 1: Markets with varying demand over items. Y-axis shows the share of agents who ranked each item as their top preferred item.

The rank minimizing (RM) mechanism finds a set of matchings $(i, j) \subset S \times M$ such that the average rank of the items to which the agents are matched is minimized. The RM mechanism is implemented using the Hungarian algorithm [5,9,11,12], by assuming some cost $c(i, j)$ for matching agent i to item j. Henceforth we refer to this as RM-HAL.

Strategies for manipulating RM-HAL are considered for two scenarios: (i) single-step cost and (ii) linear cost. In the first scenario, cost is zero for matching agents to their top-n preferred items, and constantly high for others. We analyze the steps of the Hungarian algorithm to derive a best response strategy for a strategic agent with complete information on others' preferences. The use of this best response strategy by all agents results in a Nash equilibrium.

Theorem 1. *The best response strategy for agent i to be matched to their top-n preferred items with RM-HAL is $(p_{i,1}, \ldots p_{i,n}, j_1, ..., j_{l-n})$ where $f_{j,n} > 0$, $j \notin \{p_{i,1} \ldots p_{i,n}\}$, and $|N| \geq l$.*

Deriving a best response strategy when costs are linear over rank is non-trivial, as the agent has to consider the correlations between other agents' preferences. For this scenario, a *first-order heuristic strategy* is proposed: beyond top-n items, the agent selects items with $f_{j,1} > 0$ and orders them in descending order of popularity. This heuristic aims to delay the agent being matched to items beyond the top-n, but is not guaranteed to be optimal. Deriving a best response strategy for this case is left for future work.

3 Experimental Results

A simulation study is conducted using both synthetic and real-world datasets to evaluate the impact of the first-order heuristic strategy in matches made using

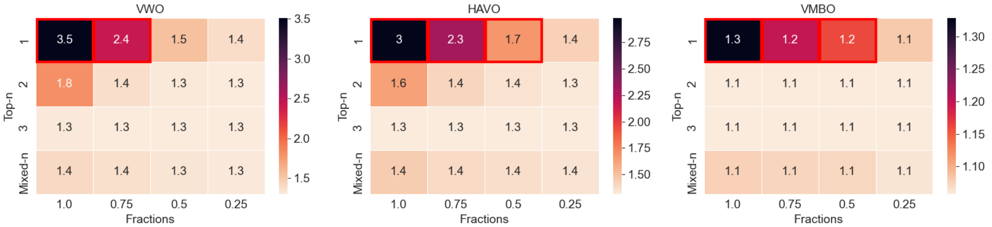


Fig. 2: Average rank of RM-HAL matches when agents apply the first-order heuristic strategy for varying values of n and f. Scenarios where RM-HAL has a worse average rank than RSD are highlighted in red.

RM-HAL. The synthetic data set models three types of markets (logistic, linear, and exponential) with varying demand distributions of 10 items between 2000 agents. The real-world dataset, sourced from the Amsterdam school choice system, contains preferences of 7,500 students across three education levels (VWO, HAVO, VMBO). The demand for the items in these datasets is visualized in Figure 1. Strategic preferences are simulated using the first-order heuristic strategy, with the assumption that agents truthfully ranked their top-n choices. The matching performance of RM-HAL using strategic preferences is compared with RSD using truthful preferences. Four experiments were conducted, varying levels of strategic manipulation, number of strategizing agents and assessing its effects on overall rank efficiency and impact on strategic and truthful agents. Strategic scenarios involved varying n (acceptable top ranks) and f (fraction of strategic agents), as well as mixed-n cases, where groups of agents used different n values.

We find that despite strategic manipulation, RM-HAL provides matches with a better average rank than RSD in most scenarios, as shown in Figure 2. Only in extreme scenarios, where more than 50% agents apply the first-order heuristic strategy to be matched to their top-1 or top-2 items, the average rank of RM-HAL is worse than RSD. We also find that applying the first-order heuristic strategy is effective but risky for agents, particularly for $n = 1$. The rank efficiency of matches is also market-dependent, with RM-HAL performing significantly better than RSD in competitive markets. However, the impact of strategy is also worse in these markets. Strategic agents also always have a better average rank than truthful students, implying that using RM-HAL in practice can lead to unequal outcomes between the two groups.

This work aims to promote discussion on the nature of strategic manipulations and their impact for efficient mechanisms such as RM-HAL. While RM-HAL is not obviously manipulable, the first-order heuristic strategy is found to be effective and easy to apply, as it only requires information on the relative popularity of items. Our simulation study shows improved match probabilities for top-n preferred items when applying the heuristic strategy. Although the strategy is not devoid of risks, these results suggest an incentive for agents to employ strategic preference reporting.

References

1. Abdulkadiroğlu, A., Pathak, P., Roth, A.: Strategy-proofness versus efficiency in matching with indifferences: Redesigning the nyc high school match. American Economic Review **99**(5), 1954–78 (2009)
2. Abdulkadiroğlu, A., Pathak, P., Roth, A., Sönmez, T.: The boston public school match. American Economic Review **95**(2), 368–371 (2005)
3. Abdulkadiroğlu, A., Sönmez, T.: Random serial dictatorship and the core from random endowments in house allocation problems. Econometrica **66**(3), 689–701 (1998)
4. Aksoy, S., Azzam, A., Coppersmith, C., Glass, J., Karaali, G., Zhao, X., Zhu, X.: School choice as a one-sided matching problem: Cardinal utilities and optimization. arXiv preprint arXiv:1304.7413 (2013)
5. Bazaraa, M., Jarvis, J., Sherali, H.: Linear Programming and Network Flows 4th ed. (2010)
6. Biró, P.: Applications of matching models under preferences (2017)
7. Bloch, F., Cantala, D.: Dynamic assignment of objects to queuing agents. American Economic Journal: Microeconomics **9**(1), 88–122 (2017)
8. Bogomolnaia, A., Moulin, H.: A new solution to the random assignment problem. Journal of Economic theory **100**(2), 295–328 (2001)
9. Burkard, R., Dell'Amico, M., Martello, S.: Assignment Problems (2009)
10. Featherstone, C.R.: Rank efficiency: Modeling a common policymaker objective. Unpublished paper, The Wharton School, University of Pennsylvania.[25, 28, 45] (2020)
11. Jonker, R., Volgenant, T.: Improving the hungarian assignment algorithm. Operations Research Letters **5**(4), 171–175 (1986)
12. Munkres, J.: Algorithms for the assignment and transportation problem. Journal of the Society for Industrial and Applied Mathematics **5**(1) (1957)
13. Ortega, J., Klein, T.: The cost of strategy-proofness in school choice. Games and Economic Behavior (2023)
14. Roth, A.: Deferred acceptance algorithms: History, theory, practice, and open questions. international Journal of game Theory **36**(3), 537–569 (2008)
15. Svensson, L.G.: Strategy-proof allocation of indivisible goods. Social Choice and Welfare **16**, 557–567 (1999)
16. Tasnim, M., Weesie, Y., Ghebreab, S., Baak, M.: Strategic manipulation of preferences in the rank minimization mechanism. Autonomous Agents and Multi-Agent Systems **38**(2), 44 (Sep 2024). https://doi.org/10.1007/s10458-024-09676-3, https://doi.org/10.1007/s10458-024-09676-3
17. Troyan, P.: Non-obvious manipulability of the rank-minimizing mechanism. arXiv preprint arXiv:2206.11359 (2022)