Sufficient Social Reasoning to Learn Productive Cycles of Cooperation

Chase McDonald

Second Year Paper
Department of Social and Decision Sciences
Carnegie Mellon University
chasemcd@cmu.edu

Abstract

Social systems are often characterized by the diverse capabilities and needs of the individuals within them. In many settings, complementary capacities and needs present an opportunity for cooperation that can increase both individual outcomes and overall social welfare; however, it may not always be the case that dyadic cooperation leads to beneficial outcomes, as any two individuals may not have the specific capabilities and needs to complement each other. That is, the system lacks a coincidence of desires. Rather, there may exist cooperative cycles in which cooperation can exist along a chain of individuals such that every individual provides and receives productive support. We characterize such situations as complex social dilemmas and investigate, using a learning-theoretic approach, the preferences and reasoning capabilities that are sufficient to enable the emergence of productive cooperation cycles. Specifically, we carry out computational experiments in the context of a multi-agent resource exchange game—the Mars Colony game—and demonstrate how distinct utility functions, which correspond to varying levels of other-regarding preferences and social reasoning, lead to emergent cooperative cycles. We draw on existing work on cooperation in social dilemmas and multi-agent learning to motivate the choice of each utility function, which corresponds to selfish and other-regarding preferences, direct and indirect reciprocity, and causal attribution. The results show that standard other-regarding preferences are insufficient in our setting. Causal attribution, in tandem with other-regarding preferences, is shown to be an effective mechanism for learning productive cooperative cycles.

Acknowledgements I acknowledge the tremendous amount of support and patience from Cleotilde Gonzalez, Simon DeDeo, and John Miller that made this paper possible—the value of their guidance cannot be overstated. This research was supported by AFRL Award FA8650-20-F-6212 sub-award number 1990692 to Cleotilde Gonzalez.

1 Introduction

Consider a setting with a group of three people, Alex, Percy, and Rue, each working on a research paper. Each of the three has their own strengths when it comes to different aspects of their projects, and each could use help in particular areas. Alex excels at data analysis, but may be bad at writing; Percy might have extensive experience in experimental design, but no experience in data analysis; and Rue could be an excellent writer but is unable to design an experiment. In this simple setting, there is a clear opportunity to increase both individual and social welfare through a cycle of support: Alex helps Percy with data analysis, Percy helps Rue with experimental design, and Rue helps Alex with writing. If the benefit of getting help is greater than the cost of providing it, such a cycle would increase both the individual and collective welfare. The cyclic support demonstrates a productive *cooperative cycle*: chains of support between agents, where the actions of one individual complement the "needs" of another.

Beyond the aforementioned example, we can consider more generally how individuals in social groups have a diverse set of capabilities and needs. The former can be used to help others, while reliance on others may be needed for the latter. In settings with sufficiently diverse capabilities and needs, dyadic cooperation may not be able to increase welfare, as any two individuals may not have complementary capabilities and needs. Instead, there may exist productive cooperative cycles, between chains of individuals, that can result in social and individual benefit.

The necessity for cycles is the result of a lack of "double coincidences of desires" (Hoshino et al., 2013; Jevons, 1876), which originates in commodity exchange scenarios with individuals or institutions with diverse sets of commodity allocations, but without the opportunity for efficient bilateral exchange. These situations have been formalized as Wicksellian exchange games (Wicksell, 2013). In everyday social interaction, centralized institutions may not exist to facilitate these cycles, and we require an alternative mechanism by which cooperative cycles can be realized. Marimon et al. (1990) discuss a Wicksellian exchange paradigm proposed by Kiyotaki and Wright (1989), where participants—in this case, autonomous agents—must *learn* strategies of exchange in a decentralized fashion. In its setting, the outcome of this process results in an argument for the use of fiat currency, which acts as a mechanism to facilitate bilateral exchange without the "double coincidence of wants." Fiat currency is not the only way in which this exchange problem can be solved. Araujo (2004) demonstrates that social norms of gift giving can act as a substitute and allow effective exchange, given that a set of population constraints are met. Indeed, Araujo demonstrates that cooperative prosocial behavior can be an effective mechanism in social exchange cycles.

In the present work, we take a learning-theoretic perspective and ask how productive cooperative cycles can emerge from the bottom up. That is, through micro-level interaction, where individuals must learn how to interact with one another by exercising their own capacities and identifying those of others. Where Araujo (2004) investigates the role of social norms in paired interaction, we investigate the sufficient social preferences and reasoning capabilities of agents that can facilitate the emergence of productive cooperation cycles in freely interacting populations of diverse agents. We employ a series of computational experiments with agents who learn via Bayesian updating. In each of the experiments, we derive a set of utility functions that correspond to distinct social preferences and reasoning capabilities and test their effectiveness in facilitating emergent cooperative cycles through simulation.

In the first experiment, we tested the efficacy of selfish versus prosocial preferences. We show that without considering the outcome of others, selfish agents are unable to learn effective behavior. In experiments 2 and 3, we introduce extensions of prosocial agents with different reasoning capabilities. In the former, we consider agents who learn other-regarding preferences based on direct and indirect reciprocity. In the latter, agents maintain other-regarding preferences, but derive utility from the impact of their own action. This is akin to asking, "did the person I help realize any benefit?" Our experiments consider both cases where agents can and cannot determine their own causal impact on other agents, which results in a refined question of "how much did *my action* help?" Our results show that, in order to overcome the difficulty of learning in a dynamic, multi-agent system, causal reasoning with prosocial preferences proves to be most effective for learning productive cooperation cycles.

Next, we review some of the motivating literature on social dilemmas, cooperation, and learning in multi-agent environments. We then present our experimental results and offer an interpretation of the results and potential implications for behavioral modeling and organizational incentive design.

Social Dilemmas & Preferences. Dawes (1980) provides the two defining criteria for a social dilemma: (1) given any set of strategies for others, individuals receive higher payoffs for defecting; and (2) each individual is better off if everyone cooperates rather than defects. Important to social dilemmas is the interdependence of each member of the social group (Van Lange & Balliet, 2015); each individual's action can influence others' experiences, which means that cooperation is the result of joint decisions. At a high level and with an appropriate reward structure, the previously described setting of cooperative cycles can be cast as a social dilemma: any individual has an incentive not to provide support to others to avoid incurring a cost, although if individuals find and engage in productive cycles, they are better off than they all decided to do nothing—to "defect".

There are several solution concepts for social dilemmas that allow groups to achieve beneficial outcomes. Dawes (1980) describes several approaches, including the alteration of payoffs and socially oriented utility functions. The former refers to structural changes in the setting that alter the task at hand, which can often be implausible. Socially-oriented utility, or motivational solutions (Kollock, 1998), offers a mechanism that avoids structural changes to the interaction setting while being capable of achieving the same goal. Socially-oriented utility functions, or those that take into account the outcomes of other relevant actors in the group, reflect levels of *social value orientation* (SVO) and can describe an intrinsic transformation of the payoff structure (Van Lange, 1999). That is to say, preferences that individuals have towards the outcomes of others offer a way to alter the payoff structure without external changes. In its simplest form, the social value orientation can be represented as a weighted sum over individual outcomes, with the relative weights on other individuals' outcomes representing the level of SVO. In this form, SVO can be used to solve some full-information social dilemmas, such as the Prisoner's Dilemma. The study of SVO, both empirically and analytically, has been widespread. Balliet et al. (2009) conduct a meta-analysis of empirical studies and show that in empirical settings, there is a significant positive relationship between cooperation and SVO in Prisoner's Dilemma games, public goods games, and tragedy of the commons games. In complementary work, which scales the complexity of tasks, McKee et al. (2020) show in computational experiments that populations of agents with increased SVO achieve better collective and individual outcomes in temporally and spatially extended public goods and tragedy of the commons games.

The widespread study of SVO and its empirical support in human cooperation motivates its use in the context of learning productive cooperation cycles. Our investigation focuses on the preferences and capacities that can enable this behavior to emerge in the population. Examination of SVO serves as a clear starting point for such an investigation. In certain cases, such as those previously mentioned, SVO can be an effective mechanism to solve social dilemmas (Bogaert et al., 2008). We investigate to what extent the effectiveness of SVO holds in the context of cooperative cycles.

An additional class of solutions to social dilemmas that has been well studied is that of strategic solutions (Kollock, 1998). Rather than maintaining preferences over the outcomes of other individuals, action selection strategies are rule-based and conditioned on the behavior of other individuals in the system. Prominent examples of such strategies are derived from *reciprocity*. One of the simplest examples of the effectiveness of reciprocity was demonstrated in the Prisoner's Dilemma tournaments by Axelrod (1984): the winning strategy in the tournament was Tit-for-Tat, which prescribes cooperation on the first round and mimicking the partners behavior in subsequent rounds. If your partner cooperates, you reciprocate with cooperation — and the same applies to defection. Reciprocity extends well beyond simple mechanisms like Tit-for-Tat, and forms of both direct and indirect reciprocity have come to represent distinct mechanisms for cooperation. The latter mechanism generalizes to larger populations and often operates on the foundation of reputation. In contrast to direct reciprocity, where an individual cooperates with someone based on the pair's previous interactions, indirect reciprocity entails each member's pro-social behavior increasing their reputation. In the indirect reciprocity setting, individuals are more likely to cooperate with those with a better reputation. This mechanism has been shown to overcome social dilemmas such as the tragedy of the commons (Milinski et al., 2002). Furthermore, Rand et al. (2014) synthesize the findings on cooperation in public goods games and describe how making actions public can facilitate empirical reciprocal behavior that leads to cooperation. In recent computational work, in the same social dilemma and public goods contexts as McKee et al. (2020), Eccles et al. (2019) demonstrate how autonomous agents that learn reciprocity strategies can induce cooperation and improve group outcomes.

Just as prior work on SVO motivates our use of other-regarding preferences, the extensive literature on reciprocity and its distinct forms does, as well. We formalize concepts in the existing literature on direct and indirect reciprocity to examine if and how these considerations can support learning in the present context.

As previously described, we use both SVO and reciprocity in the context of learning. There exists much work on learning in social dilemmas, in which various preferences or capacities are studied—or prescribed, in the case of computational experiments. Macy and Flache (2002) demonstrates cases in which cooperative equilibrium can emerge with self-regarding agents based on the reward structure in two-player repeated social dilemmas, Gonzalez et al. (2015) demonstrated how human-like cooperative behavior can be learned via dynamic adjustments to an agent's SVO, and the previously mentioned Eccles et al. (2019) and McKee et al. (2020) study emergent cooperation in complex social dilemmas with reciprocity and SVO, respectively. We build on the ideas on previous work by incorporating such mechanisms into the learning of cooperative cycles; a context where agents need to learn not only that cooperation can increase their welfare but also coordinate their cooperation such that it facilitates productive cycles. This requires that agents not only be

capable of learning prosocial behavior but also be able to discern effective behavior in a setting with many interacting agents.

Multi-Agent Learning. The act of learning from experience is often complicated as it is: individual agents must explore the space of possible actions and accurately update their beliefs about the world to improve their decision-making. Learning becomes increasingly difficult as the number of adapting agents increases. Each agent must learn not only how their actions impact the world, but also their joint actions with other agents who are also learning and changing their behavior. This is known as the non-stationarity problem in the multi-agent reinforcement learning literature (Canese et al., 2021; Nowe et al., 2012). Non-stationarity caused by other agents makes the credit assignment problem (Sutton & Barto, 2018) exponentially more difficult. That is, how does each agent separate the effect of their own action from that of other agents'? There are a number of approaches that attempt to minimize the impacts of non-stationarity (Chang et al., 2003); however, one line of inquiry is of particular relevance to the present work: causal reasoning. Indeed, mechanisms for causal reasoning can allow agents to marginalize the effect of other behaviors in their environment and learn the impacts of their own behavior (Foerster et al., 2017; Grimbly et al., 2021).

We are particularly concerned with the reasoning abilities of agents that are sufficient to learn cooperative behavior. As such, we must pay particular attention to problems that plague learning more generally. The literature on causal reasoning in multi-agent systems motivates our use utility functions that discern causal impact, which enables us to ask if causal reasoning is a sufficient capability for learning productive cooperation cycles. In addition to deriving utility functions that correspond to social preferences and reciprocity, we do so for causal reasoning. This allows us to ask how important causal attribution is to achieve cooperative outcomes.

In the following sections, we formalize our methods and present our experimental results.

2 Methods & Measures

In the current work, we establish a game with a population of agents with diverse skillsets such that the criteria for a social dilemma are satisfied and cycles of cooperation lead to the social welfare-maximizing outcome. We endow populations of agents with distinct utility functions, each corresponding to varying levels of other-regarding preferences and reasoning capabilities, then simulate learning in the game to study the emergence of complex cooperative strategies. The setting of the game represents an extreme case where individual capabilities are stringently defined, and utility functions take a homogeneous form within a population. Nonetheless, the simulations offer an account of specific capabilities that can lead to cooperative behavior.

2.1 Mars Colony Game

In this work, we present an n-player resource exchange game: the Mars Colony game. In this setting, there is a set $\mathcal{A}$ of n agents and an initial allocation of resources for each agent. Throughout the game, agents must learn to transfer resources between each other to maximize their utility, which is defined as a function of the change in the allocation of resources. Each agent $i \in \mathcal{A}$ is associated with a particular conversion

capability $\mathcal{C}_i(\cdot)$. For each resource $r \in \mathcal{R}$, $\mathcal{C}_i : \{r\} \to \{\eta r'\}$, where $\eta \geq 1$ is a scalar multiplier, and r' is any resource in the resource set $\mathcal{R}$. Agents use their conversion abilities to increase their resources, but can also choose strategies to transfer a number of its own resources to another agent. Each agent i maintains an inventory $\mathcal{I}_i$ that describes its current allocation, and an agent can only transfer a resource r to another agent if it is present in its inventory, that is, $r \in \mathcal{I}_i$. All agents have a policy $\pi_i(\mathcal{I}_i)$ that determines their strategy. The policy is conditioned on the current set of resources available in each agent's inventory. An action $a^{(t)} = \{j, r\}$ details the agent receiving a resource ($j \in \mathcal{A}$) and the resource being transferred (r). Importantly, agents can also take a "no-op" action, where they do nothing, that is, $a^{(t)} = \emptyset$.

Given transfer and conversion abilities, the overall welfare —- in terms of the number of resources —- of a group of agents is maximized when transfers exist that facilitate efficient conversion cycles. For instance, for a group of three agents $i \in \{1, 2, 3\}$ who have conversion abilities $\mathcal{C}_1(\{A\}) = \{3B\}$, $\mathcal{C}_2(\{B\}) = \{3C\}$, and $\mathcal{C}_3(\{C\}) = \{3A\}$, the first should transfer B to the second, who in turn should transfer C to the third, and finally the third transfers A back to the first. This type of behavior constitutes an efficient trading cycle—a productive cooperation cycle—in the population, where agents maximize the total number of resources created in the system.

The full game procedure is described by Algorithm 1.

Algorithm 1 Mars Colony Game Procedure

Require: $\mathcal{A}, T \in \mathbb{R}^+, \mathcal{I}_i \forall i \in \mathcal{A}$ ▷ Initialize the set of agents $\mathcal{A}$, timesteps T, and inventories $\mathcal{I}$.
  $t \leftarrow 1$
  $\mathcal{I}_i \leftarrow C_i(\mathcal{I}_i) \, \forall i \in \mathcal{A}$ ▷ Possible conversions are made before transfers
  while $t \leq T$ **do**
    $A^{(t)} \leftarrow \{a_i \sim \pi_i(\mathcal{I}_i) \, \forall i \in \mathcal{A}\}$ ▷ All agents simultaneously select actions
    for $(j, p) \in A^{(t)}$ **do** ▷ Each action specifies a transfer of resource p to agent j
      $\mathcal{I}_j \leftarrow \mathcal{I}_j \cup \{p\}$
    end for
    $\mathcal{I}_i \leftarrow C_i(\mathcal{I}_i) \, \forall i \in \mathcal{A}$ ▷ All agents make possible conversions.
    $t \leftarrow t + 1$
  end while

The Mars Game constitutes a social dilemma, as it satisfies the two criteria described by Dawes (1980): for any agent, if they choose to take no action ("no-op"), its own utility is strictly greater than it would be if, all other strategies equal, they transferred any resource to any other player. Furthermore, if all agents choose strategies such that they transfer resources to agents that can make positive conversions, each agent's utility is strictly higher than if they took no action; that is, if they "defected". There is a critical distinction between this setting and standard social dilemmas: not all cooperative behavior is productive. Indeed, the dynamics become significantly more complex than scenarios with binary cooperate or defect strategies; rather, the effectiveness of a cooperative strategy is dictated by the strategies, capabilities, and needs of all agents collectively.

An important difference of the Mars Colony game, when compared to many n-player cooperation settings, is that players are not randomly paired in their interactions (as in, e.g., Marimon et al. (1990)). Instead, each agent can choose who to interact with and this interaction need not be reciprocal. This is to say that at every timestep, all agents select both *who* to transfer a resource to (if any) as well as which resource to transfer.

2.2 Bayesian Learning

We consider agents who learn via Bayesian updating. In particular, the agents use maximum *a posteriori* updates to learn a distribution over utilities x for each possible action. For each action k, agent i estimates the parameters $\theta_{i,k}$ of a sampling distribution $f(\mathcal{D}|\theta_{i,k})$, where $\mathcal{D} = \{x_i\}_{i=1}^n$ is a set of n observed draws from the true distribution—the observed data. Given a prior distribution $g(\theta_{i,k})$ over $\theta_{i,k}$, we can obtain an estimate via

$$\hat{\theta}_{i,k}(\mathcal{D}) \propto f(\mathcal{D}|\theta_{i,k})g(\theta_{i,k}) \tag{1}$$

In the previously described context, and as will be expounded upon in Section 2.3, each agent i has an internal function $\mathcal{U}_i(\cdot)$ that maps states of the world $s \in \mathcal{S}$ to their realized utility. Realized utilities are what comprise the dataset $\mathcal{D}$ which is used to estimate the parameters of the true distribution. Specifically, each agent maintains a memory of outcomes associated with each action. For all elapsed timesteps $t \in \{1, ..., T\}$

$$\mathcal{D}_{i,k} = \left\{ \sum_{j=0} \gamma^j \mathcal{U}_i(s_{t+i}) \mid a_{i,t} = k \right\}_{t=1}^{T} \tag{2}$$

where γ is a discount rate.

In our experiments, we assume a prior distribution of $\theta_{i,k} \sim \Gamma(\alpha, \beta)$ where $\alpha, \beta > 0$. This implies a probability density function of the form

$$g(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-\beta x} \beta^{\alpha}}{(\alpha - 1)!} \tag{3}$$

Further, we assume the sampling distribution takes the form of an exponential distribution, for which $\Gamma(\alpha, \beta)$ is a conjugate prior. The posterior distribution becomes $\Gamma(\alpha + N, \beta + \sum_j^N x_j)$ where N is the number of observations in $\mathcal{D}_{i,k}$ and $x_i \in \mathcal{D}_{i,k}$.

Each agent computes the expected value of each action independent of the current state of the world, which corresponds to $X_{i,k} \sim \Gamma(\alpha + N, \beta + \sum_j^N x_j)$:

$$\kappa_{i,k} = \mathbb{E}\left[X_{i,k}\right] \tag{4}$$

Then, actions are selected probabilistically, where the probability of agent i taking action k at time t is given by

$$p(a_i^{(t)} = k) = \frac{\exp\left(\kappa_{i,k}/\tau\right)}{\sum_j \exp\left(\kappa_{i,j}/\tau\right)} \tag{5}$$

where $\tau > 0$ is the temperature. As $\tau \to 0$, action selection becomes greedy and as $\tau \to \infty$, all actions become equally likely.

Critically, we note the support of $\Gamma(\cdot, \cdot)$ is given by $(0, \infty)$. In the Mars Colony Game, it is possible to incur costs: reward may be below 0. As such, we simply add a constant $C + \epsilon$ to all rewards, such that we shift the distribution of outcomes to be within the support. In practice, we set $C = n$; that is, the lowest possible reward occurs when all agents lose a resource, so offsetting by the number of agents ensures reward will always be greater than or equal to 0—and $\epsilon = 1e - 4$ to ensure inequality. In practice, rewards will never

fall to the level of $-n$; however, it is sufficient to find any constant greater than or equal to the minimum reward.

2.3 Preferences & Utility Functions

Central to our work is the specification of each agent's utility function. We consider several different forms, each of which correspond to varying levels of other-regarding preferences and reasoning capabilities. Each specific utility function is introduced in turn; however, there are several common concepts and pieces of notation that are used throughout.

First, the utility function of agent i is denoted by $\mathcal{U}_i(\cdot)$. This is a function that takes as input information about the state of the world, and returns an intrinsic reward signal to the agents.

Second, we define an operator that denotes the change over successive timesteps in resource allocations. Specifically the change in the number of resources in agent i's inventory from $t-1$ to t is given by Equation 6.

$$\Delta_i^{(t)} = |\mathcal{I}_i^{(t)}| - |\mathcal{I}_i^{(t-1)}| \tag{6}$$

The changes in resource allocations are the primary factor that serves as input to the utility functions. We represent the state at time t by $s_t = \Delta^{(t)}$, a vector of the inventory changes over time.

Lastly, we introduce a weighting matrix W, where $w_{i,j}$ is the weight of agent i's other-regarding preferences towards agent j. In our utility functions, self- and other-regarding preferences are represented as a linear combination over utilities.

In each of the experiments carried out below, we introduce utility functions that correspond to selfish agents, fully other-regarding agents, direct and indirect reciprocity, and causal attribution. In addition to their behavioral implications, they correspond to levels of social reasoning and carry with them implications about cognitive abilities, e.g., the capacity for social awareness or the ability to conduct causal inference.

2.4 Measures

The primary-dependent measure in the following simulations is *efficiency*. Efficiency measures the proportion of possible "production" currently being carried out in the population. If every agent is traded a resource for which they can make their highest-rate conversion, the population is at maximum efficiency. Formally, the efficiency is given by Equation 7.

$$\text{Efficiency}(A^{(t)}) = \frac{\sum_{\{i,r\} \in A^{(t)}} \left[|\mathcal{I}_i \cup \mathcal{C}_i(\{r\})| - |\mathcal{I}_i \cup \{r\}| \right]}{\sum_{i \in \mathcal{A}} \max_r \left[|\mathcal{I}_i \cup \mathcal{C}_i(\{r\})| - |\mathcal{I}_i \cup \{r\}| \right]} \tag{7}$$

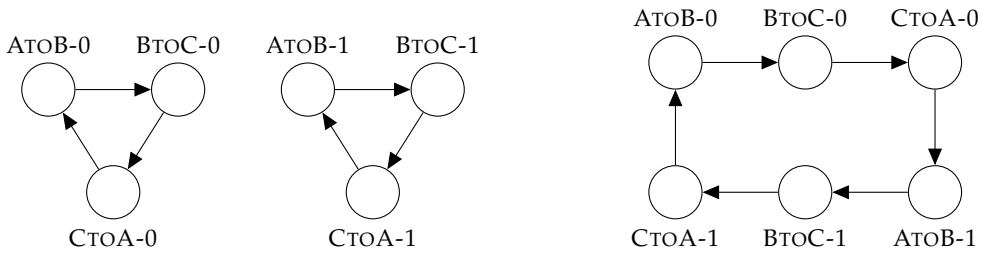
It is important to note that, in this constrained setting, agents can only convert a single resource at each timestep. So, if agent i can convert a unit of r to d units of another resource, $|\mathcal{C}_i(\{r\})| = d$ only for the first conversion of each timestep—and 1 otherwise.

3 Learning Cooperative Cycles

In this section, we conduct a series of experiments to investigate how agents with varying preferences and cognitive abilities learn productive cooperative cycles in the Mars Colony Game. We carry out three variations of each experiment corresponding to $n \in \{3, 6, 9\}$; that is, with population sizes that allow for variable numbers of emergent trade cycles. For each simulation we set the number of total timesteps to $T = 200$ and average results over 50 independent trials.

We choose populations that are multiples of three for a specific reason: there are three distinct resources in this instantiation of the Mars Colony Game: $\mathcal{R} = \{A, B, C\}$, and $\frac{n}{3}$ agents have each of one of the following conversion abilities: $C_i(\{A\}) = \{3B\}$, $C_i(\{B\}) = \{3C\}$, and $C_i(\{C\}) = \{3A\}$. Note that set notation is used here because an agent may have additional resources in their inventory, for which the conversion function C_i acts as an identity function and returns them in the output set. Each agent's inventory is initialized with three of the resources that they convert *to*. For example, if agent i has conversion ability $C_i(\{A\}) = \{3B\}$, then its inventory is initialized as $\mathcal{I}_i = \{3B\}$.

Given the described setting, we can consider the idealized outcome. That is, maximum efficiency is reached when cycles of transfer emerge in which an agent who makes positive conversions of A to B transfers B to another agent, who in turn converts B to C, and then transfers C to an agent who converts it back to A. The cycle is complete when the final agent transfers A back to an agent who can convert A to B. There are several ways for this to emerge, and we provide the two simplest cases in the context of $n = 6$ agents in Figure 1: a large cycle or two smaller cycles. As the number of agents increases, the number of possible efficient—as well as inefficient—configurations grows. This means that it is not only of interest to find settings in which cooperative cycles are learned, but also the structures that they result in. Indeed, some may be more robust to individual agent deviations than others, as is the case with the two small cycles in Figure 1a when compared to the larger cycle in Figure 1b.



(a) (b)

Figure 1: Examples of cooperative cycles with maximum efficiency. Nodes represent agents, with their labels denoting the conversion ability. For instance, AtoB-0 could convert resource A to three units of B. We assume that the agents in these figures are transferring the resource that they convert to (following the previous example, AtoB-0 would transfer resource B to BtoC-0).

The parameters for each different type of agent are fixed in the current simulations and described by Table 1. It is important to note that we use a discount rate of $\gamma = 0$, which implies that the agents do not consider the future returns. Indeed, in exploratory experiments we found that higher discount rates, in general, lead to worse performance. A presentation and discussion of these results is given in Appendix A.1.

Lastly, in order to encourage exploration of the action space, we utilize an optimistic prior (Sutton & Barto, 2018). That is, we initialize $\mathcal{D}_{i,k} \forall i, k$ with the value $2n$. This encourages agents to attempt each action in the action space and to not commit to a strategy too early in the learning process (this value is higher than any possible observed reward).

α	β	τ	γ
1.0	3.0	0.2	0.0

Table 1: Fixed parameters used across experiments.

3.1 Experiment 1: Selfish vs. Other Regarding Preferences

We first investigate the capacity of both selfish and other-regarding agents to learn strategies that lead to productive cooperative cycles in the Mars Colony game. The selfish case represents agents with an atomized utility function. They only consider the changes in their own inventory. In the case of other-regarding preferences, there is a non-zero weight on the inventory changes of other agents. We define both preferences with the following function, with selfish being the special case where $w_{i,i} > 0$ and $w_{i,j} = 0 \, \forall j \neq i$:

$$\mathcal{U}_i(s_t) = \underbrace{w_{i,i}\Delta_i^{(t)}}_{\text{Own Utility}} + \underbrace{\sum_{j \in \mathcal{A}, j \neq i} w_{i,j}\Delta_j^{(t)}}_{\text{Other-Regarding Utility}} \tag{8}$$

Here agent i derives utility from the weighted sum of changes in inventory sizes; that is, the gain or loss of resources of each agent. Importantly, other-regarding utility in this case isn't recursive: there isn't a consideration of *other* agents' pro-social preferences, or any other calculation that may enter utility functions other than the change in their inventory.

Further, we note that in preliminary experiments, we found that efficiency is monotonically increasing as a function of other-regarding weight. This motivated the choice of using what we refer to as equal weighting preferences, where $w_{i,j} = 1 \forall i, j \in \mathcal{A}$. This contrasts the selfish case, where $w_{i,i} = 1$ and $w_{i,j} = 0 \forall i \neq j$.

Intuitively, one might expect that, with enough data and an appropriate discount rate, selfish agents would *eventually* be able to learn; however, the empirical literature on pro-social preferences in social dilemmas suggests that the equal weighting case should be able to learn more effectively than the selfish agents. We also expect this to be the case because the equal weighting utility function rewards agents with a signal that is directly proportional to the efficiency of the population: a collective-welfare maximizing strategy set is identical to an efficiency-maximizing one.

Results & Discussion The results in terms of efficiency, for both selfish and other-regarding agents, are shown in Figure 2. In the $n = 3$ case, we observe the most distinctive results: the strategies of the selfish agents decrease in efficiency over time, implying that agents are learning non-cooperative strategies; on the other hand, populations of other-regarding agents reach maximum efficiency over a relatively short timespan. In the $n = 6$ and $n = 9$ cases, we see similar results; however, the effects have been attenuated: equal weighting agents fail to reach maximum efficiency, both plateauing near 50%, and the selfish populations still diminish in overall efficiency, albeit at a slower rate.

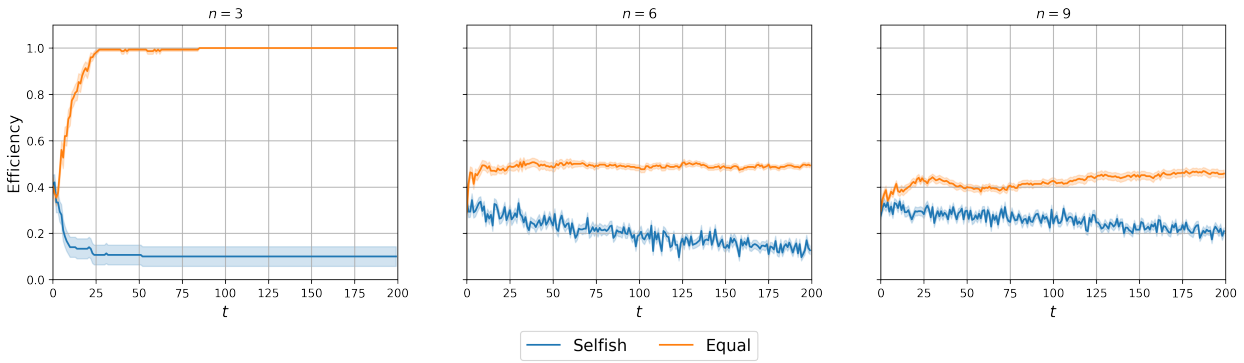


Figure 2: Efficiency for populations of agents with selfish and equal weighting ($w_{ii} = w_{ij}\,\forall i, j \in \mathcal{A}$) utility functions. The shaded regions are the standard error.

These results raise several questions. First, what strategies do the selfish agents learn such that they become less cooperative over time? The answer to this question is simple: they learn to take no action at all. Figure 3 depicts the proportion of actions taken by both groups of agents that are "no-op" actions. Rather than finding the rewarding "signal in the noise" over the course of their experience, they learn via the single discernible signal: transferring a resource results in a consistent loss and should always be avoided.

Second, we ask why equal weighting agents are unable to learn successful cooperative cycles as the population increases. If the agents receive a reward signal that is proportional to efficiency, why are they unable to learn efficient trading cycles? The answer to this question lies in the complexities of multi-agent learning: there is high variance in observed outcomes and non-stationarity throughout the group's strategies. Each agent must learn to maximize the reward signal; however, they are unable to discern the impact that their own action had on the realization of that signal. This leads to inefficient, flawed learning: some agents may be excluded from cycles or multiple agents may learn to transfer to the same recipient. The initial spike in the proportion of no-op actions, as shown in Figure 3, depicts initial exclusions: these agents aren't receiving any resources and their inventory is quickly depleted. This exclusion is in part due to the determinism of strategies: with $\tau = 0.2$, agents commit to strategies relatively quickly. The effect of altering τ, and optimizing all parameters, is explored in Appendix B.

Both the equal weighting and selfish agents suffer from similar issues, despite their gaps in performance. Selfish agents are unable to find consistent positive signals to encourage trade. Equal weighting agents are unable to consistently determine whose action led to positive signals, and thus settle at inefficient strategies.

3.2 Experiment 2: Direct vs. Indirect Reciprocity

In this section, we perform an additional experiment that introduces agents with variations of preferences for direct and indirect reciprocity. Direct and indirect reciprocity has been shown as a mechanism to encourage cooperative behavior, and often solves defection issues in social dilemmas with repeated interactions (Komorita et al., 1991; Okada, 2020).

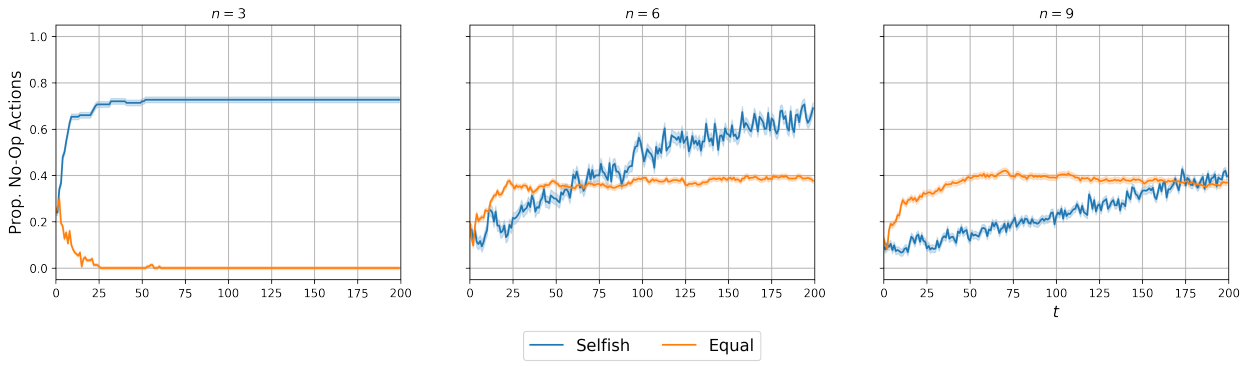


Figure 3: The proportion of no-op actions taken as a function of time for selfish agents versus equal weighting agents. In the selfish case, agents learn abstain from transferring resources. In the $n = 6$ and $n = 9$ cases, the proportion of no-op actions increases because the agents have failed to learn efficient strategies, and some agents run out of resources to transfer: they are left out of trading cycles and are forced to no-op.

We focus on reciprocity in terms of levels of *reciprocal preferences*; each agent's utility weights the gain or loss of other agents higher or lower as a function of their contributions to the current agent. Formally, we define two kinds of reciprocal preferences: standard and recursive. Both rely on a dynamically changing weight matrix W. That is,

$$w_{i,j} = \begin{cases} \frac{R_{i,j}}{\sum_k R_{i,k}} & \text{if } i \neq j \\ 0.5 & \text{otherwise} \end{cases} \tag{9}$$

where $R_{i,j}$ is the number of resources agent i has received from agent j. Importantly, this quantity represents post-conversion resources: if agent i can convert resource A to three units of B, and agent j gives one unit of A to agent i, $R_{i,j}$ will increment by three. Standard reciprocal preferences are then given by the same form as in Equation 8 with the altered W matrix. It is also important to note that we alter the weight on own utility to be 0.5: this enables the possibility of prosocial strategy learning, as $R_{i,j} \leq 1$, the weight on other agents would be unlikely to reach equal weighting, which means a greedy agent would never act to transfer an (unconvertible) resource to another agent.

Recursive reciprocal preferences, on the other hand, are given by

$$\mathcal{U}_i^{RR}(s_t) = w_{i,i}\Delta_i^{(t)} + \sum_{j \in \mathcal{A}, j \neq i} w_{i,j} \left[\Delta_j^{(t)} + \underbrace{\sum_{k \in \mathcal{A}} w_{j,k}\Delta_k^{(t)}}_{\text{Recursive Reciprocal Utility}} \right] \tag{10}$$

In simple terms, recursive preferences extend the utility function to consider the utility of individuals who give to agent i and those who give to them: I care about people who are nice to me and those who are nice to the people that are nice to me. This implies that an agent's weight on their own inventory change is itself adapting. We can isolate the terms in Equation 10 corresponding to agent i such that the incorporation of their own utility becomes $\Delta_i^{(t)} \left[w_{i,i} + \sum_{j \in \mathcal{A}, j \neq i} w_{i,j} w_{j,i} \right]$.

There are natural generalizations of recursive reciprocal preferences to what we will call level-k recursive reciprocation. The selfish case can be thought of as $k = 0$, the standard reciprocation case as $k = 1$, and the

setting in Equation 10 as $k = 2$. In this sense, standard reciprocal preferences represent direct reciprocity and recursive preferences represent indirect reciprocity—and increasingly so as $k \to n$.

The motivation for employing such utility functions is to refine the information that agents are receiving. These act as a middle-ground between the selfish and equal weighting utility functions described in Section 3.1: they allow for selective social values that encourage supporting other agents in ways that might facilitate cooperative cycles. Such utility functions restrict the noise that is introduced in the equal weighting case—where the results of all agent actions enter at once—while still incorporating social information that the selfish agents lack.

Results & Discussion The results in terms of efficiency shown over time are presented in Figure 6. The results of equal weighting agents from Section 3.1 are included for comparison.

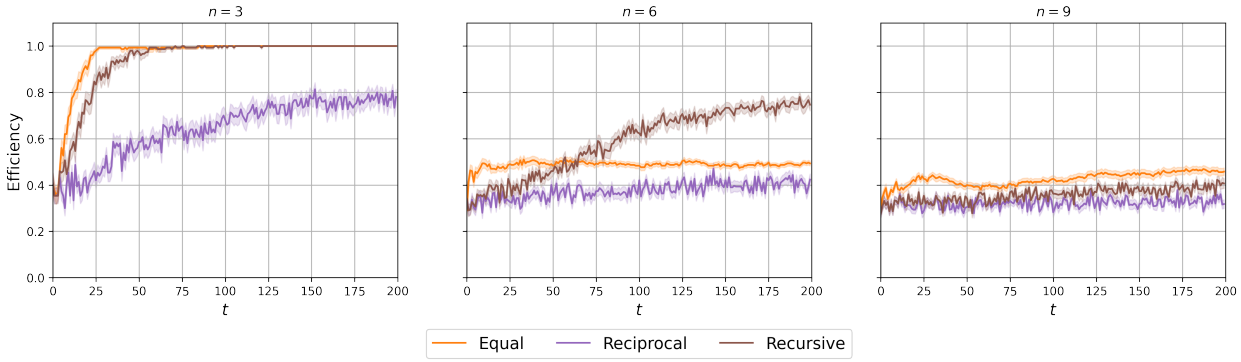


Figure 4: Efficiency for populations of agents with reciprocal and recursive utility functions, compared against equal weighting.

For reciprocal and recursive preferences, agents display learning in the $n = 3$ case. Agents with reciprocal preferences converge to roughly 80% of maximum efficiency, whereas the agents with recursive preferences are able to learn fully cooperative strategies and reach maximum efficiency. In both the $n = 6$ and $n = 9$ cases, the former fails to learn sufficiently cooperative strategies, only marginally increasing efficiency over the course of the simulation. However, in the $n = 6$ case, the recursive preferences enable agents to surpass the efficiency of equal weighting agents, but fail to do so in the $n = 9$ case, where they only exhibit marginal learning in terms of efficiency.

As in the previous setting, we investigate why exactly reciprocal and recursive preferences fail or succeed in facilitating emergent cooperative cycles. First, consider the implications of direct reciprocity from the reciprocal utility function: agents are incentivized to learn strategies that result in resources being transferred to other agents that give directly to them. There are two basic ways in which this can manifest. First, the agent may learn that the agent who gave to them is best off when they receive resources that they can convert; thus, it should learn to transfer resources in a way that increases the probability of the other agent receiving a convertible resource. This would lead to efficient cycles. Alternatively, it may be the case that such cycles are too difficult to learn due to the signals being too noisy. In this setting, agents may increase the probability that they reciprocate directly (e.g., transfer resources directly to agents that transfer to them) or to take no action at all; indeed, it may be the case that the learned weights on other agents' gains are

not high enough to make transfer beneficial. The results shown here depict the latter. Agents are typically unable to learn in the $n = 6$ and $n = 9$ cases that transferring to an independent third agent can increase the utility more than directly transferring to an agent that transfers to them.

Second, we investigate the success of recursive preferences in the $n = 3$ and $n = 6$ settings. We observe significant improvements in efficiency by incorporating the recursive reciprocation; indeed, indirect reciprocity presents itself as a potential solution when the population is relatively small. In the $n = 3$ setting, there is sufficient incentive to close the cooperative cycle. In the $n = 6$ case, the signal persists; however, if initial learning results in larger cycles, as depicted in Figure 1b, agents won't have sufficient incentive to close the cycle. Such a setting, and its implications, are demonstrated in Figure 5.

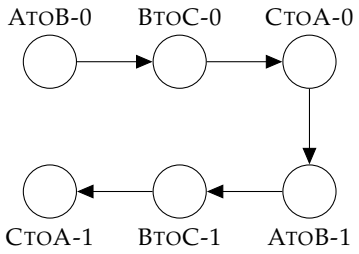


Figure 5: An illustrative failure case for a population of $n = 6$ agents with utility functions that reward recursive reciprocation, as defined by Equation 10. No outgoing edge represents the no-op action. Here, CTOA-1 will fail to close the cycle if ATOB-0 has not transferred them a sufficient number of resources in the past. Similarly, they will not make a transfer to ATOB-1 until enough transfers have occurred via ATOB-1 $\rightarrow$ BTOC-1 $\rightarrow$ CTOA-1. These cycles are not sustainable, as ATOB-0 will run out of resources to transfer; as the weights change, so will the network structure, but there is no guarantee that the cycle will close to be the desired form.

Lastly, in the $n = 9$ case, the previous difficulties are accentuated: there is an increased number of opportunities for failure cases, similar to those in Figure 5, and the increased non-stationarity caused by the larger group learning makes it increasingly difficult to discern positive signals for learning.

3.3 Experiment 3: Causal Attribution Preferences

In this experiment, we introduce a final pair of utility functions that are meant to alleviate the difficulty in learning with many actors: preferences for an agents own impact on other agents. The first, naive approach at formalizing such preferences is through directed preferences: utility is derived from the gain or loss in resources of the agent that is traded to. Specifically, given an action

$$a_t = \begin{cases} \{j, p\} & \text{if resource } p \text{ was transferred to agent } j \\ \emptyset & \text{otherwise} \end{cases} \tag{11}$$

the utility for agent i is then given by

$$\mathcal{U}_i^D(s_t|a_t) = w_{i,i}\Delta_i^{(t)} + \sum_{j \in \mathcal{A}, j \neq i} \delta[j \in a_t] w_{i,j} \Delta_j^{(t)} \tag{12}$$

where $\delta[j \in a_t]$ is an indicator function that is 1 if $j \in a_t$ and 0 otherwise. That is, agent i considers only the change in their own allocation and the allocation of the agent that they transfer a resource to. The directed utility function directly rewards agents for transfers that result in successful conversions; however, it does not represent true causal inference. There may be other confounding factors that influence the receiving agent's inventory: they gave a resource away, or another agent transferred to them.

The final mechanism we consider is one in which agents are able to discern the true causal impact that their action had on other agents. In particular, denote by $\Delta_i^{(t)}(a)$ the change in agent i's inventory at time step t that was the result of action a. For example, if agent k takes action $a = \{i, p\}$—that is, transferring resource p to agent i—and agent i can convert p to d units of another resource, we would have that $\Delta_i^{(t)}(\{i, p\}) = d$. This causal attribution mechanism is akin to a refined version of directed preferences, as $\Delta_i^{(t)}(\{j, p\}) = 0$ for any $j \neq i$ and for all p. The utility function takes the form of

$$\mathcal{U}^{CA}(s_t|a_t) = w_{i,i}\Delta_i^{(t)} + \sum_{j \in \mathcal{A}, j \neq i} w_{i,j}\Delta_j^{(t)}(a_t) \tag{13}$$

where W can take any of the previously described forms (fixed or adaptive). In the simulations for both directed and causal utility functions, we use a fixed W where every entry is 1.

The mechanisms for causal attribution overcomes one of the most difficult problems in multi-agent learning: credit assignment. Indeed, in each of the preceding utility functions, a key difficulty in learning has been on discerning the impact of the agent's own action. For instance, in the equal weighting case, when agents are making initial choices and altering their strategies jointly, there is no clear signal as to which agent led to an ultimate increase or decrease in the observed outcome. Here, we remedy that issue by assuming causal reasoning capabilities and significantly reduce the variance in utility signals.

Results & Discussion The learning curves for efficiency are presented in Figure 6. Directed preferences lead to roughly 80% efficiency in the $n = 3$ case, but diminish in efficiency for both the $n = 6$ and $n = 9$. On the other hand, causal preferences result in maximum efficiency in the $n = 3$ case, and surpass the equal weighting agents in both the $n = 6$ and $n = 9$ settings.

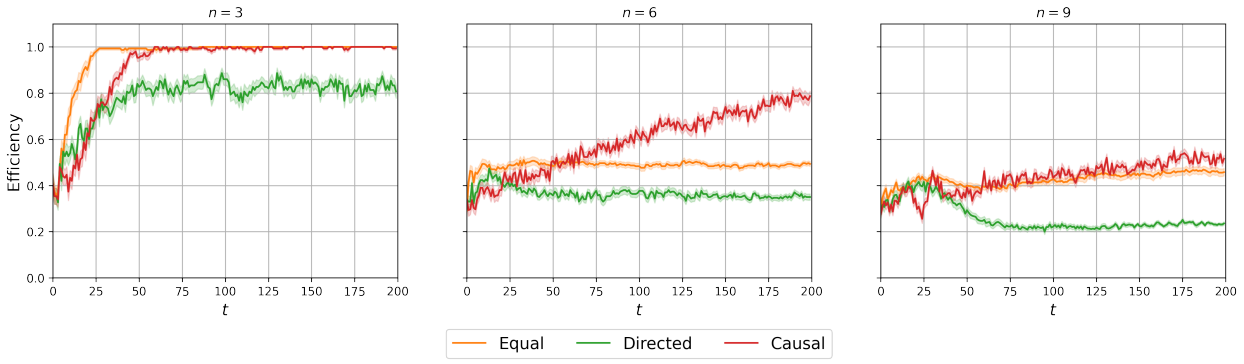


Figure 6: Efficiency for populations of agents with directed and causal utility functions, compared against equal weighting.

Figure 7: A comparison of success and failure cases in the directed preferences setting. In Figure 7a, AtoB-0 and CtoA-0 are both transferring resources to BtoC-0, which means that they observe the net impact of both of their transfers. Presuming that they transfer resource B and A, respectively, and BtoC-0 makes no transfer, both of the former agents observe $\mathcal{U}^D(s^{(t)}|a^{(t)}) = 4$. If CtoA-0 changes its strategy to transfer A to AtoB-0, both agents will only end up receiving $\mathcal{U}^D(s^{(t)}|a^{(t)}) = 3$. This emphasized by the successful cycle in Figure 7b, where each agent receives a utility of 2; there is incentive to (temporarily) deviate for a higher directed utility.

The failure of the directed utility agents in learning cooperative strategies is explained by the noisy signal that they receive about their actions. Consider the following scenario with three agents, depicted graphically in Figure 7. If two agents both transfer a resource to a third agent, and one of them is a resource that the third agent can make a positive conversion of, they will both observe a utility that is higher than what they would get if both make efficient transfers to separate agents. Furthermore, as the population grows to $n = 6$ and $n = 9$, the probability that multiple agents transfer to the same other agent increases, which results in a decrease in efficiency.

The causal preferences setting eliminates this issue. When agents are endowed with the capacity to conduct causal reasoning—and have preferences that reward their causal contributions to others—they are able to discern scenarios in which multiple agents are transferring to the same recipient and alter their strategies accordingly.

3.4 Experiment 4: Curse of Dimensionality & Institutional Constraints

Our final experiment aims to address a structural issue that makes learning difficult: the "curse of dimensionality" in the action space. Indeed, as the number of agents grows, we increase the number of possible strategies for an agent. In the $n = 9$ case, there 25 possible actions that any agent can take, and they must have sufficient experience with each in order to discern its effectiveness. This is compared to the 7 possible actions in the $n = 3$ case. In social worlds, individuals or institutions may have information about the capabilities or needs of others. Consider the opening example again: one individual who is a good data analyst may know that someone else needs help with data analysis, and would benefit from their support. They do not need to consider all other individuals, only the relevant subset. Such an idea motivates this simulation: establishing an "institutional constraint" whereby individuals can only transfer resources to other agents that can make positive conversions. For example, resource B can only be transferred to agents who can convert B into more than one unit of another resource.

We follow the same procedure as described in the previous experiments, utilizing each of the previously described utility functions.

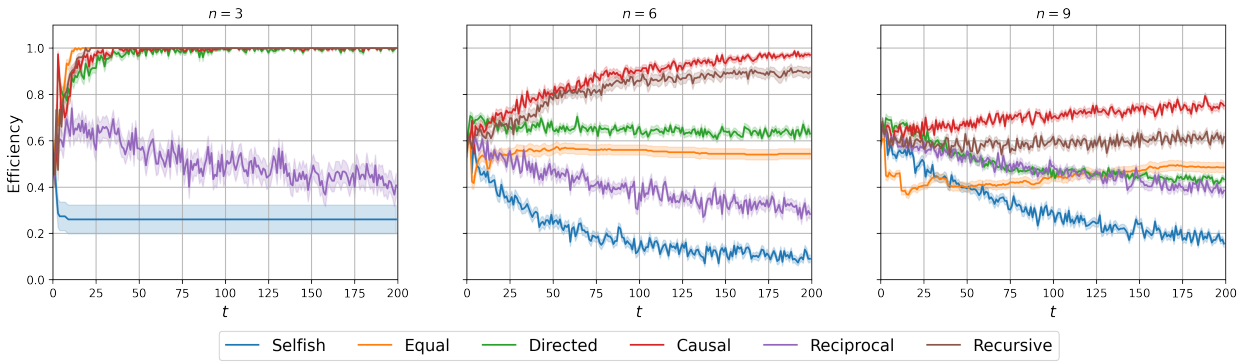


Figure 8: Efficiency for populations of agents with each utility function; however, transfers here are restricted such that any resource can only be transferred to an agent who can make a positive conversion.

Results & Discussion Efficiency over time in the institutional constraint setting is shown in Figure 8. All utility functions except reciprocal and selfish are able to converge to maximum efficiency in the $n = 3$ case; in this simplified setting, which dramatically restricts the action space to efficient transfers, learning is only inhibited by a lack of prosocial preferences and task incongruent preferences: that is, agents learn to value transfers in the reciprocal case that would not be permitted with the institutional constraint. For example, if AtoB-0 transfers B to BtoC-0, the weight of the former's gain increases in the latters utility function, but as a result of the imposed constraint, BtoC-0 cannot transfer the resource it has produced (C) back to AtoB-0. More generally, the results in the $n = 3$ setting with the institutional constraint emphasize that the bar for sufficient social reasoning abilities and other-regarding preferences is low in small groups.

As the group size increases, the different utility functions differentiate themselves: both causal and recursive preferences lead to near-optimal efficiency over the course of learning, and selfish and reciprocal preferences both diminish in efficiency over time for the same reasons as previously described. Both equal weighting and directed preferences remain relatively consistent over time: in the directed case, agents suffer from the same failure cases as described in Figure 7. In the equal weighting setting, individuals settle on inefficient strategies because of the noisy equal weighting signal—variance in utility signals and non-stationarity prevent learning with equal-weighting, despite the scaled down action space. These patterns are consistent for both the $n = 6$ and $n = 9$ settings, although directed preferences lead to a diminished efficiency in the $n = 9$ setting, as the previously described difficulties are again exasperated. Despite relative improvements from the unconstrained setting, adding the institutional constraint failed to alleviate many of the issues that inhibit learning of cooperative cycles.

4 General Discussion

The results presented here highlight the types of reward signals that can lead to the emergence of productive cooperative cycles in social dilemmas, and in doing so emphasize the social reasoning capabilities necessary for agents to learn in such settings. We demonstrated the inability of selfish agents to learn cooperative strategies in the standard Mars Colony game. In such social settings, with many individuals learning, selfish signals do not provide sufficient information for agents to learn that they can be better off by incurring an immediate cost. By introducing equal weighting agents, we demonstrate how pro-social

preferences can facilitate the emergence of cooperative cycles. As the population increases beyond a single cycle, generic equal weighting alone proves to be insufficient for learning productive cycles.

In an explorative series of experiments to find the utility information that is sufficient the learning of these cooperative cycles, we demonstrated the efficacy of variations of direct and indirect reciprocity, and two forms of causal attribution. For the former, we showed that direct reciprocity is insufficient in the small-group setting; however, a form of indirect reciprocity—as facilitated by recursive preferences—can offer a significant improvement. Lastly, across group sizes, we demonstrate that the consideration of an agent's causal contribution to other agent welfare is the best-performing mechanism in terms of emergent cooperation; however, it is important that causal reasoning is precise: directed preferences that don't discern true causal impacts can be detrimental to social welfare.

These results have a number of implications in relatively distinct domains. We highlight two of them here: in cognitive modeling in multi-agent systems and in organization behavior. For the former, our results highlight the need for methods of causal inference when building models of humans—or any autonomous agents—that must reason in a social world. We highlight an idealized setting and, rather than reason about *how* exactly individuals are able to make causal judgments, we endow them with that ability and simulate the results. There are a number of efforts in multi-agent learning to develop methods for discerning causal impact (see, e.g., Foerster et al. (2017) and Grimbly et al. (2021)). The findings we have presented here demonstrate the potential efficacy of accurate causal reasoning mechanisms as it pertains to cooperative learning, and underscore the importance of incorporating such capabilities into computational models.

To the latter point of organization behavior, our results suggest that there are particular mechanisms for incentive design that can facilitate the emergence of cooperative cycles in groups and, as a result, group efficacy. In particular, our results suggests that incentives that take the form of causal impact contain sufficient information about individuals in order for cooperative cycles to be learned. This is particularly important in ad hoc groups or those without prior knowledge of one another: when behavioral strategies must be learned, group size and reward scheme play a key role in how cooperative behavior may emerge.

Future Work There are many possible ways to extend the analysis presented here. We describe several direct avenues for future work that can build on our simulation results.

Individual Differences Preferences and abilities undoubtedly vary from person to person. Here, we have considered a simplified case of homogeneous preferences. There are a myraid of different ways in which heterogeneity may be introduced into the framework proposed here, and future work may consider the sufficient proportions or combinations of agents that can result in learned cooperative cycles. For instance, the strength of other-regarding preferences can be varied in a population—as is typical in empirical scenarios (Bogaert et al., 2008)—and one might ask how characteristics of the distribution of such preferences results in emergent cooperation. Similarly, the reasoning capacities may be varied—selfish agents may be in a population of agents with causal preferences—and one can answer questions analogous to those previously discussed.

Network Implications As briefly discussed in Section 3, there are a number of ways in which cooperative cycles may manifest: from a single large cycle on one extreme to a larger number of smaller cycle. The structure of the emergent network may, in some cases, have significant impacts and the ways in which dif-

ferent kinds of information or utility functions impact network formation would be important. Continued analysis may investigate the implications on the utility functions for emergent networks and the ways in which utility functions can be designed to couple cooperative learning with a desired network structure.

Human Subjects Experiments An experiment with human subjects would also be insightful. This would test how humans learn in the same setting and serve to validate the present findings. In a follow-up empirical experiment, we plan to test how humans in the Mars Colony Game learn when we dictate the type of reward signal they receive. For instance, we can manipulate how study participants are rewarded based on their, e.g., own allotment of resources versus their cumulative causal effect on others. In a similar vein, alternative studies can be carried out that provide all information to human participants as they make their decisions, and then a post-hoc analysis can be conducted to see how human behavior aligns with the simulation results. This may answer the question of what kind of utility function particular individuals are using to learn cooperative cycles in these tasks.

A Experiment Variations

A.1 Selfish Agents and γ

We perform a number of variations on the selfish-agent experiment in Section 3.1, varying the discount rate γ to demonstrate an inability for agents to learn productive cooperative cycles for any time horizon. The results for $\gamma \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ are shown in Figure 9.

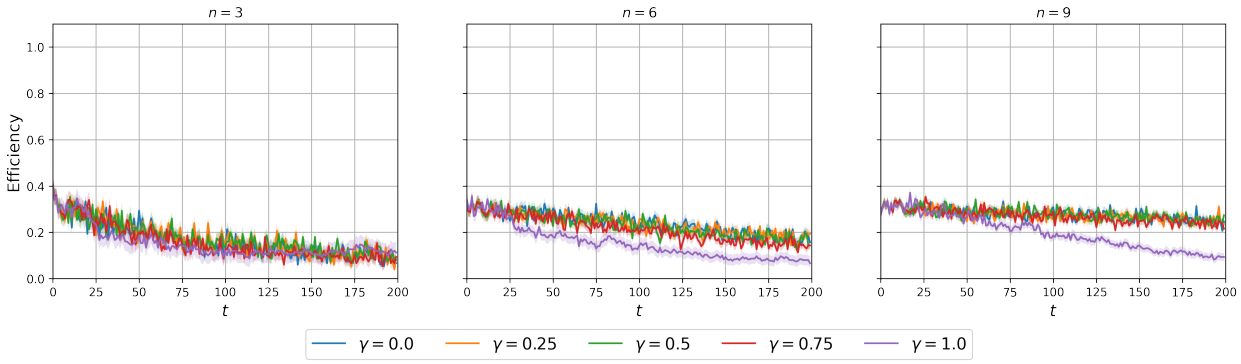


Figure 9: Efficiency for selfish agents ($w_{ij} = 0 \forall i, j \in \mathcal{A}, i \neq j$) for different values of γ.

B Robustness

In Section 3.1, we fix the parameters of all agents to allow for easy comparison between each. Here, we offer an alternative view where we conduct a search over the parameter space for each preference type, then use the parameters that maximize the average efficiency. The results are shown in Figure 10 and the resulting parameters are presented in Table 2.

There are several changes in relative performance that are worthwhile to consider. One notable change is that directed preferences demonstrate some improvements in efficiency in the $n = 6$ and $n = 9$ settings.

Furthermore, recursive preferences are no longer superior to equal weighting. Lastly, the selfish agents no longer decrease over time in efficiency. This is a result of the parameters being selected to make the selfish agent effectively random. Table 2 shows that the parameter for τ is the maximum in our search space (recall that as $\tau \to \infty$, each action becomes equally likely, regardless of observed outcomes).

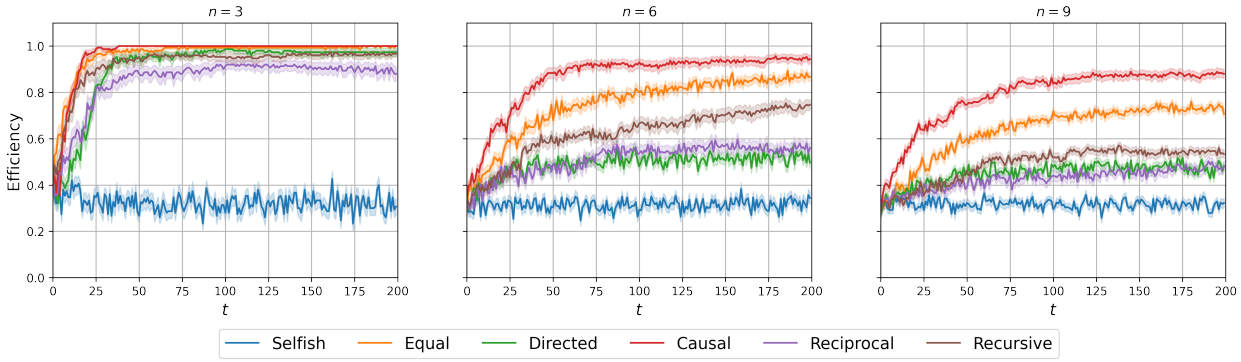


Figure 10: Efficiency for agents after conducting a sweep over the parameter space.

References

Araujo, L. (2004). Social norms and money. *Journal of Monetary Economics*, *51*(2), 241–256. https://doi.org/10.1016/j.jmoneco.2003.01.005

Axelrod, R. M. (1984). *The Evolution of Cooperation* [Google-Books-ID: NJZBCGbNs98C]. Basic Books.

Balliet, D., Parks, C., & Joireman, J. (2009). Social Value Orientation and Cooperation in Social Dilemmas: A Meta-Analysis. *Group Processes & Intergroup Relations*, *12*(4), 533–547. https://doi.org/10.1177/1368430209105040

Bogaert, S., Boone, C., & Declerck, C. (2008). Social value orientation and cooperation in social dilemmas: A review and conceptual model. *The British Journal of Social Psychology*, *47*(Pt 3), 453–480. https://doi.org/10.1348/014466607X244970

Canese, L., Cardarilli, G. C., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M., & Spanò, S. (2021). Multi-Agent Reinforcement Learning: A Review of Challenges and Applications [Number: 11 Publisher: Multidisciplinary Digital Publishing Institute]. *Applied Sciences*, *11*(11), 4948. https://doi.org/10.3390/app11114948

Chang, Y.-h., Ho, T., & Kaelbling, L. (2003). All learning is Local: Multi-agent Learning in Global Reward Games. *Advances in Neural Information Processing Systems*, *16*. Retrieved April 20, 2022, from https://proceedings.neurips.cc/paper/2003/hash/c8067ad1937f728f51288b3eb986afaa-Abstract.html

Dawes, R. M. (1980). Social Dilemmas. *Annual Review of Psychology*, *31*(1), 169–193. https://doi.org/10.1146/annurev.ps.31.020180.001125

Eccles, T., Hughes, E., Kramár, J., Wheelwright, S., & Leibo, J. Z. (2019). Learning Reciprocity in Complex Sequential Social Dilemmas [arXiv: 1903.08082]. *arXiv:1903.08082 [cs]*. Retrieved April 20, 2022, from http://arxiv.org/abs/1903.08082

	Tuned				Fixed	
	α	β	τ	γ	$w_{i,i}$	$w_{i,j}$
Selfish	4.0	1.0	2.0	0.2	1.0	0.0
Equal	2.0	4.0	0.2	0.0	1.0	1.0
Reciprocal	1.0	2.0	0.0	0.3	0.5	Adaptive
Recursive	2.0	3.0	0.2	0.0	0.5 (Adaptive)	Adaptive
Directed	0.0	1.0	0.0	0.2	1	1
Causal	2.0	0.0	0.001	0.0	1.0	1.0

(a) $n = 3$

	Tuned				Fixed	
	α	β	τ	γ	$w_{i,i}$	$w_{i,j}$
Selfish	3.0	5.0	2.0	0.1	1.0	0.0
Equal	1.0	4.0	0.6	0.0	1.0	1.0
Reciprocal	1.0	1.0	0.08	0.2	0.5	Adaptive
Recursive	1.0	1.0	0.15	0.1	0.5 (Adaptive)	Adaptive
Directed	1.0	5.0	1.4	0.4	1.0	1.0
Causal	4.0	5.0	0.001	0.0	1.0	1.0

(b) $n = 6$

	Tuned				Fixed	
	α	β	τ	γ	$w_{i,i}$	$w_{i,j}$
Selfish	0.001	3.0	2.0	0.7	1.0	0.0
Equal	1.0	4.0	0.8	0.0	1.0	1.0
Reciprocal	1.0	1.0	0.08	0.2	0.5	Adaptive
Recursive	1.0	1.0	0.15	0.1	0.5 (Adaptive)	Adaptive
Directed	1.0	5.0	1.4	0.4	1	1
Causal	4.0	5.0	0.001	0.0	1.0	1.0

(c) $n = 9$

Table 2: Parameter results after performing a grid search over the parameter space.

Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2017). Counterfactual Multi-Agent Policy Gradients [arXiv: 1705.08926]. *arXiv:1705.08926 [cs]*. Retrieved July 4, 2021, from http://arxiv.org/abs/1705.08926

Gonzalez, C., Ben-Asher, N., Martin, J. M., & Dutt, V. (2015). A Cognitive Model of Dynamic Cooperation With Varied Interdependency Information. *Cognitive Science*, *39*(3), 457–495. https://doi.org/10.1111/cogs.12170

Grimbly, S. J., Shock, J., & Pretorius, A. (2021). Causal Multi-Agent Reinforcement Learning: Review and Open Problems [arXiv: 2111.06721]. *arXiv:2111.06721 [cs, stat]*. Retrieved April 22, 2022, from http://arxiv.org/abs/2111.06721

Hoshino, Y., Ishikawa, R., & Yamazaki, A. (2013). *Unequal Distribution of Powers in a Wicksellian Transfer Game* (tech. rep. No. 24). Meisei University, School of Economics. Retrieved April 26, 2022, from https://ideas.repec.org/p/mei/wpaper/24.html

Jevons, W. S. (1876). *Money and the Mechanism of Exchange* [Google-Books-ID: D3QqAAAAYAAJ]. D. Appleton.

Kiyotaki, N., & Wright, R. (1989). On Money as a Medium of Exchange. *Journal of Political Economy*, *97*(4), 927–954. Retrieved April 26, 2022, from https://www.jstor.org/stable/1832197

Kollock, P. (1998). Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, *24*(1), 183–214. https://doi.org/10.1146/annurev.soc.24.1.183

Komorita, S. S., Hilty, J. A., & Parks, C. D. (1991). Reciprocity and Cooperation in Social Dilemmas. *Journal of Conflict Resolution*, *35*(3), 494–518. https://doi.org/10.1177/0022002791035003005

Macy, M. W., & Flache, A. (2002). Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, *99*(suppl_3), 7229–7236. https://doi.org/10.1073/pnas.092080099

Marimon, R., McGrattan, E., & Sargent, T. J. (1990). Money as a medium of exchange in an economy with artificially intelligent agents. *Journal of Economic Dynamics and Control*, *14*(2), 329–373. https://doi.org/10.1016/0165-1889(90)90025-C

McKee, K. R., Gemp, I., McWilliams, B., Duéñez-Guzmán, E. A., Hughes, E., & Leibo, J. Z. (2020). Social diversity and social preferences in mixed-motive reinforcement learning [arXiv: 2002.02325]. *arXiv:2002.02325 [cs]*. Retrieved April 26, 2022, from http://arxiv.org/abs/2002.02325

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002). Reputation helps solve the 'tragedy of the commons' [Number: 6870 Publisher: Nature Publishing Group]. *Nature*, *415*(6870), 424–426. https://doi.org/10.1038/415424a

Nowe, A., Vrancx, P., & De Hauwere, Y.-M. (2012). Game Theory and Multi-agent Reinforcement Learning. *Adaptation, Learning, and Optimization* (p. 30). https://doi.org/10.1007/978-3-642-27645-3_14

Okada, I. (2020). A Review of Theoretical Studies on Indirect Reciprocity. *Games*, *11*(3), 27. https://doi.org/10.3390/g11030027

Rand, D. G., Yoeli, E., & Hoffman, M. (2014). Harnessing Reciprocity to Promote Cooperation and the Provisioning of Public Goods. *Policy Insights from the Behavioral and Brain Sciences*, *1*(1), 263–269. https://doi.org/10.1177/2372732214548426

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (F. Bach, Ed.; 2nd ed.). A Bradford Book.

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*(2), 337–349. https://doi.org/10.1037/0022-3514.77.2.337

Van Lange, P. A. M., & Balliet, D. (2015). Interdependence theory. *APA handbook of personality and social psychology, Volume 3: Interpersonal relations* (pp. 65–92). American Psychological Association. https://doi.org/10.1037/14344-003

Wicksell, K. (2013). *Lectures on Political Economy (Routledge Revivals): Two Volumes* [Google-Books-ID: Cf4MsWwHst0C]. Routledge.