Emergent Resource Exchange and Tolerated Theft Behavior Using Multiagent Reinforcement Learning

Jack Garbus*
Brandeis University
Department of Computer Science
garbus@brandeis.edu

Jordan Pollack
Brandeis University
Department of Computer Science

Abstract For decades, the evolution of cooperation has piqued interest in numerous academic disciplines, such as game theory, economics, biology, and computer science. In this work, we demonstrate the emergence of a novel and effective resource exchange protocol formed by dropping and picking up resources in a foraging environment. This form of cooperation is made possible by the introduction of a campfire, which adds an extended period of congregation and downtime for agents to explore otherwise unlikely interactions. We find that the agents learn to avoid getting cheated by their exchange partners, but not always from a third party. We also observe the emergence of behavior analogous to tolerated theft, despite the lack of any punishment, combat, or larceny mechanism in the environment.

Keywords
Emergence, multiagent, reinforcement learning, cooperation, behavior

1 Introduction

Although many consider human intelligence a core factor in the success of our species, perhaps an even more fundamental component is our ability to cooperate with one another (Henrich, 2016). It is no surprise, then, that the emergence of cooperative behavior has long been an area of interdisciplinary study as researchers seek to model societal dynamics (Epstein & Axtell, 1996; Gostoli & Silverman, 2023) and design artificial intelligence (AI) to work with humans (Bhoopchand et al., 2022; Gauthier & Mordatch, 2016).

Axelrod (1984) demonstrated the power of cooperation through ecological simulations where a diverse population of strategies faced each other in the Iterated Prisoner's Dilemma (IPD) game. Cooperative strategies that punished instances of defection eventually dominated the population, while defecting strategies proved to be unstable in the long run. The resulting theory from these experiments has been broadly applied to understand the emergence of cooperation between organisms, such as bacteria as well as enemy troops during wartime (Axelrod, 1984).

The IPD has since cemented itself as a vital tool for understanding the evolution of cooperation through the use of evolutionary algorithms (EAs), which allow strategies to mutate and evolve over time (Axelrod, 1987). EAs have shown that in both error-prone and error-free versions of IPD, populations in cooperative equilibrium can be invaded by mutants due to drift and the evolution of increasingly complex strategies (García & van Veelen, 2018; Lindgren, 1992). In other domains

* Corresponding author.

  https://doi.org/10.1162/artl_a_00423

where cooperation is not an elementary action in the environment, EAs can evolve complex cooperative strategies constructed from the actions they possess (Burtsev & Turchin, 2006).

In recent years, deep multiagent reinforcement learning (MARL) has emerged as another viable tool for studying cooperative behavior (Agapiou et al., 2023; Hughes et al., 2018; Leibo, Perolat et al., 2019; McKee et al., 2021). Whereas EAs perform selection and mutation on large populations of individuals to maximize fitness, MARL algorithms perform backpropagation on a smaller set of neural networks to maximize the reward obtained in an environment. Many MARL environments have been developed in pursuit of a variety of emergent social behaviors, such as turn taking, teaching, resource sharing, reciprocity, and language (Agapiou et al., 2023; Gupta et al., 2021; Lazaridou et al., 2017). The success of multiagent reinforcement learning has even led some researchers to outline a path toward artificial general intelligence heavily oriented around reproducing human social intelligence in silico (Leibo, Hughes et al., 2019). If we view the success of humanity as a story of cooperation rather than isolated intelligence (Henrich, 2016), then this research path may be promising.

Despite the potential of social intelligence, research on artificial societies has remained limited. Although agent-based models allow researchers to study changes to social behavior, they often employ fixed, human-designed approximations of real-world dynamics and behaviors (Epstein & Axtell, 1996; Gostoli & Silverman, 2023; Hinsch & Bijak, 2023). Large language models have recently enabled alternative, flexible forms of social modeling, such as *social simulacra*, which simulate community interactions between different personas provided by prompts (Park et al., 2022, 2023). These large models are trained on vast amounts of human-generated data, which plays a large role in the behaviors of the social model, thus limiting the study of how intelligent or optimal behaviors may first arise. Systems that optimize agent behavior using reinforcement learning are typically configured to study a specific set of emergent social behaviors and often utilize additional modifications to the algorithm or environment's mechanics. Some additions include the training of additional classifiers alongside each policy (Vinitsky et al., 2022), auxiliary losses (Bhoopchand et al., 2022), or behavior-specific mechanisms to enable desired behavior, such as trading (Johanson et al., 2022; Suarez & Isola, 2022). If we are to realize the vision of emergent artificial societies, we would like to discover simple, general-purpose environmental mechanisms that can induce different social behaviors instead of adding an additional layer of complexity per behavior.

To this end, we demonstrate the emergence of resource exchange without programming any form of exchange protocol into the algorithm or environment. Although agents have successfully leveraged human-designed exchange systems and discovered how to barter, trade stocks, and devise tax policies (learning to game the subsequent tax system as well) (Johanson et al., 2022; Pricope, 2021; Zheng et al., 2020), no prior work to our knowledge has demonstrated emergent resource exchange by picking up and placing resources in an embodied setting, despite the apparent simplicity of the behavior. We call our environment the Trading Game, as agents discover the ability to trade resources by picking up and placing down foraged resources.

Food sharing is a prevalent phenomenon among various species, and its evolution has been a topic of interest for researchers in evolutionary biology and anthropology (Kaplan et al., 1985). Kaplan et al. reviewed several hypotheses that have been proposed to explain resource sharing from an evolutionary perspective in nonhuman species, including kin selection, tolerated theft, reciprocal altruism, and cooperative acquisition. The kin selection hypothesis suggests that sharing resources with close genetic relatives enhances the fitness of shared genes. The tolerated theft hypothesis proposes that individuals with ample resources allow those without to steal from them, as defending the resource is more costly than the resource's value. The reciprocal altruism hypothesis, previously applied to the IPD, suggests that reciprocating cooperative behavior can emerge and be an evolutionarily stable strategy, and the cooperative acquisition hypothesis proposes that social carnivores hunt together to increase their chances of catching prey. In the Trading Game described herein in which agents cannot fight, hunt, or reproduce, we see that reciprocal altruism is the primary driver stabilizing the emergence of exchange. In addition, we observe the emergence of tolerated theft, despite the lack of any method for stealing resources or engaging in combat.

Our contributions are as follows:

- We introduce the Trading Game, a simple foraging environment that applies pressure for agents to congregate around a campfire at night.
- We demonstrate that agents in our environment can learn to exchange resources using drop/pickup actions during a period of extended congregation, whereas agents from prior research environments could not.
- We demonstrate the emergence of a behavior akin to tolerated theft between agents in our environment, despite the lack of any combat, punishment, or larceny system.
- Through an ablation study, we demonstrate how reciprocated resource exchange fails to emerge without sufficient pressure to congregate for extended periods of time.

2 Background

In this section, we briefly review a few of the most relevant environments used to study emergent cooperation and embodied exchange.

2.1 Cleanup

The Cleanup environment poses a complex social dilemma for multiagent research and has been used to study how systems for reputation, social influence, inequity aversion, and public sanctioning shape emergent cooperation (Hughes et al., 2018; Jaques et al., 2019; McKee et al., 2021; Vinitsky et al., 2022). In Cleanup, agents must simultaneously clear pollution from a river and collect apples, which spawn proportionally to the amount of pollution cleared. To prevent free-riders from collecting apples without clearing pollution, agents are equipped with a "punish beam," which they can fire at other agents to fine them with a certain amount of negative reward. This punishment mechanism allows agents to punish free-riders that do not contribute to the cleaning effort. When the punishment beam is combined with one of the additional systems mentioned earlier agents learn to punish free-riders to achieve socially beneficial outcomes and overcome the social dilemma of free-riders.

2.2 Fruit Market

In the Fruit Market environment described by Johanson et al. (2022), agents can move around, produce and consume apples or bananas, and broadcast one of 19 human-designed trade offers to nearby agents that are automatically executed by the environment once an agent accepts. Agents are designated as Apple Farmers (producing more apples at a time) or Banana Farmers (producing more bananas at a time). When Apple Farmers receive a larger reward for consuming bananas than apples, they are incentivized to produce apples and trade them for bananas (and vice versa for Banana Farmers). Agents eventually converge on trading as the optimal strategy, and a behavior akin to bartering soon emerges, where agents broadcast the offer that most benefits them, lowering their prices when they encounter other agents who do the same with counter offers. Agents adjust their offers until an agreement is reached, after which the transaction is executed by the environment. When agents were given the ability to drop and pick up items as an alternative to hand-engineered offers, agents learned to avoid freely giving away resources, thus necessitating the implementation of a trading mechanism for exchange to occur. Further experimentation that varied the relative supply and demand of resources showed corresponding changes in prices akin to what one might expect from real-world supply–demand curves. It was also shown that under certain maps with apples and bananas on opposite sides, a "merchant"-like behavior can emerge, where agents trade for apples on the apple-saturated side and then sell them at a higher price on the banana-saturated side.

2.3 AI Economist

Aside from Fruit Market, other environments directly implement different mechanisms for exchange. Of note is the AI Economist detailed by Zheng et al. (2020), which tasks agents with gathering wood and stone to construct houses. Agents are set with different skill multipliers such that some agents are able to gather more resources, while others make more coins building houses. Additionally, the environment provides a global market to which agents can submit buy and sell orders from anywhere on the map, which are automatically executed once a valid transaction exists. This environment adds an additional 44 actions for trading alone, representing the combination of 11 different price levels, whether the order is a buy or a sell, and whether the resource is wood or stone. While adding substantial environmental complexity, the market enables agents to specialize in gathering or building houses and to trade for the materials or coins they need.

2.4 NeuralMMO

The exchange system of NeuralMMO described by Suarez and Isola (2022) also introduces a global market through which agents can buy and sell items using gold. Alongside the exchange system, a profession system is introduced that allows agents to produce items needed by other professions. As a result, agents from each profession must purchase items from other professions to progress and raise their level. Agents can sell items by posting them to the market along with a price, and agents can simply select an existing item on the marketplace to purchase it at the specified cost. The NeuralMMO exchange system introduces 161 unique item types, making it one of the most complex exchange systems in a multiagent research environment.

3 Method

3.1 Multiagent Reinforcement Learning

Our environment is represented as a partially observable Markov decision process described by the tuple $< S, O, A, P, R, \gamma, N >$. The observation function O maps each state $s \in S$ to the local observation o_t^i of the environment at time step t for agent i. The shared action space between N agents is denoted by A. Each agent is controlled by a policy $\pi(\theta)$ that is parameterized by a weight vector θ. In this setting, agents act one at a time rather than simultaneously. The probability transition function $P(s'|s_t^i, a_t^i)$ is represented by P, where a_t^i is the action taken by agent i at time step t and s' is the new environment state after the action has been taken. Notably, $s' = s_t^{i+1}$ if other agents still need to take their turn for the current time step; otherwise, $s' = s_{t+1}^1$. The reward function is denoted by $R(s_t, a_t)$, and the discount parameter is represented by γ. The objective of each agent i is to maximize its discounted accumulated reward over an episode of T time steps, which is represented by $\mathbb{E}_{a_t^i, s_t^i}[\sum_{t=0}^{T} \gamma^t R(s_t^i, a_t^i)]$.

3.2 Environment

We formulate our environment as a two-dimensional grid world with two types of resources, fruits and greens, denoted by red and green squares, respectively. Five fruits spawn randomly in each of the two patches in the left corners of the grid, and five greens spawn randomly in each of the two patches in the right corners. Agents are able to move up, down, left, or right or perform no action. Additionally, agents can pick all fruits or greens on their cell and can place 0.5 fruits or 0.5 greens on their cell, resulting in nine total actions.

Agents automatically consume 0.1 units of whatever resources they possess on every time step. If an agent can consume only one type of food, it does not fulfill all its nutritional needs and receives 0.1 reward. If an agent consumes both a fruit and a green in a single step, then its needs are fulfilled, and the agent receives a reward of 1. Thus, to maximize reward, agents should consume fruits and greens together. Unlike in Johanson et al. (2022), all agents share the same reward function and are equally proficient at resource collection. Agents also receive an additional "collection" reward equal

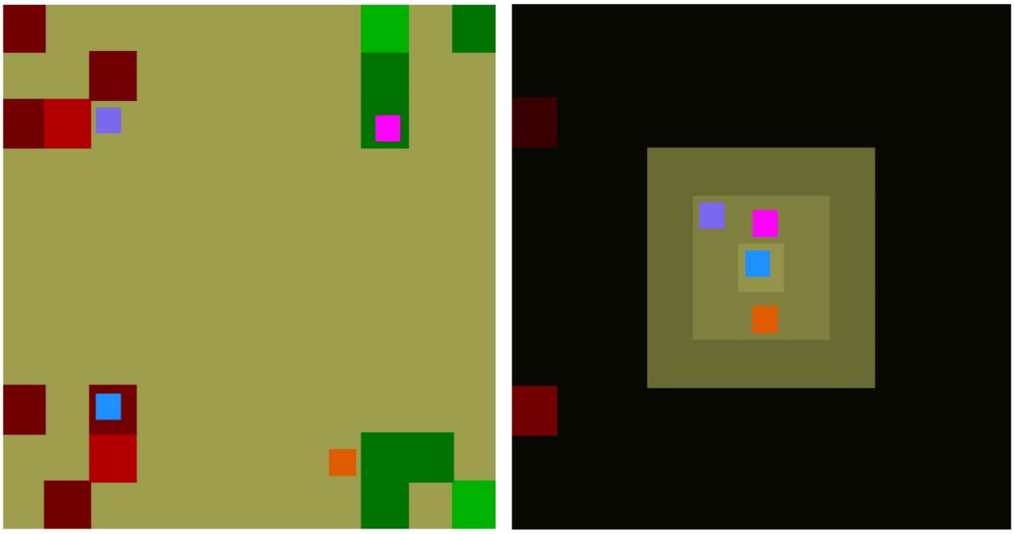

Figure 1. (left) The Trading Game environment during the daytime. Four agents are denoted in purple, pink, blue, and orange, collecting fruits and greens from patches located in the corners of the grid. Brighter red and green cells indicate that more resources are located in those cells. (right) The Trading Game environment during the night time. All four agents are located around a campfire, which provides enough light to avoid the night time penalty in a 3 × 3 area. The outer 5 × 5 ring of the fire provides enough light to reduce, but not negate, the light penalty.

to the number of newly spawned resources they collected that time step. Resources placed by other agents do not contribute to the collection reward, as it would be possible to generate large amounts of reward by repeatedly placing and picking up resources.

Unlike many environments used to study the emergence of cooperation in multiagent systems, our environment has a day/night cycle (Figure 1). The light level l for each cell on the grid starts at 0, which is the start of a new day, and then increases to 1 before oscillating between −1 and 1 for the rest of the episode in small steps. To incentivize agents to avoid dark regions of the map, we introduce a light penalty p, which is set to 10 by default. Agents lose lp reward when on cells with a negative light level, which scales the punishment by the darkness of the cell. With the default value of $p = 10$, agents can lose up to 10 units of reward in a single step during the middle of the night.

For agents to survive the darkness without receiving continual punishment, there is a small "campfire" region in the center of the grid that produces a faint light in a 5 × 5 area. The internal 3 × 3 area around it holds a light level greater than zero throughout the entire episode, and the outer ring holds a light level just under zero. The addition of the day/night cycle adds an element of periodicity to our setting; instead of wandering around the entire episode collecting food, we can expect agent behavior to alternate between foraging during the daytime and joining up around the campfire at night, treating the campfire like a "home base," as described by Isaac (1978). Agents begin each episode in one of the four corners of the campfire's 3 × 3 area, spawning in the same corner each episode. At the start of each day, all remaining resources on the grid are removed, and two patches of new fruits and greens spawn randomly around the four corners of the map.

Days and nights each last 24 time steps, which is enough time for agents to acquire all the resources in a single patch during the day. For an agent to acquire both types of resources on its own, it must stay out during a portion of the night. All experiments shown last 180 time steps, which simulates 4 days of foraging, terminating at midnight on the fourth day. For the purposes of our analysis, we only report exchange statistics from the first three nights, as agents do not always finish trading when the episode terminates in the middle of the fourth night.

Table 1. Observation channels.

Channel	Description
Fruits	No. of fruit on a grid cell
Greens	No. of greens on a grid cell
Light level	Light level for each cell
Self position	1 where the agent is located
Self fruits	No. fruit carried by the agent
Self greens	No. greens carried by the agent on cell
Policy 1 position	No. agents controlled by policy 1 on cell
Policy 1 fruits	No. fruit carried by agents controlled by policy 1
Policy 1 greens	No. greens carried by agents controlled by policy 1
…	…
Policy 4 position	1 for cells containing an agent controlled by policy 4
Policy 4 fruits	No. fruit carried by agents controlled by policy 4
Policy 4 greens	No. greens carried by agents controlled by policy 4

Note. Values for each channel are zero where the description does not specify otherwise. Fruits, greens, and position channels have values only on cells where an agent is controlled by the respective policy. Notably, because the environment allows agents to share the same policy, we provide a set of "self" channels to differentiate an agent from others sharing the same policy. For the experiments shown, this is not needed, as our agents all use different policies.

3.3 Observations

Agents observe a local 7×7 area of the grid centered around themselves. For our setting with four agents, each cell contains 18 channels of data, yielding observation tensors of shape $(7, 7, 18)$. A description of each channel can be found in Table 1. As agents act sequentially, each observation contains the state of the environment after the previous agent has acted.

3.4 Algorithm

We train a deep neural network as our policy using the proximal policy optimization (PPO) algorithm (Schulman et al., 2017), leveraging the implementation provided in the Ray Python library[1]. Although many algorithms are tailored for multiagent settings (Lowe et al., 2020; Rashid et al., 2018), vanilla PPO has been shown to be effective on many multiagent problems (de Witt et al., 2020; Yu et al., 2021). All of our agents utilize separate policies that share no parameters.

Each policy contains a vision network, a memory layer, and a controller network, controlled by a convolutional network, a long short-term memory (LSTM) layer (Hochreiter & Schmidhuber, 1997), and a multilayer perceptron, respectively. The output of the vision network is fed into the memory layer, and the output of the memory layer is sent to the controller, which produces action

1 https://github.com/ray-project/ray.

Table 2. Policy architecture.

Submodule	Layer	Description
Vision network	2-D convolution	128 3 × 3 filters, padding = 1, stride = 1
	2-D convolution	128 3 × 3 filters, padding = 1, stride = 1
	2-D convolution	128 3 × 3 filters, padding = 1, stride = 1
	Flatten	
Memory network	LSTM	512 hidden units
Controller network	Dense	256 hidden units
	Dense	256 hidden units

probabilities. The architecture details can be found in Table 2, and the full list of hyperparameters is available in the appendix.

4 Results

Our primary setting uses a heavy light penalty, $p = 10$, and five resource units per patch. For this setting, we ran 10 trials in total but only discuss results from 9 trials in sections 4, 4.2, and 4.3, as one of the trials converges to a unique minimum, which we cover separately in section 4.1.

As seen in Figure 2, the behavior throughout training in the primary setting can be viewed as two periods of equilibrium with a period of fluctuation in between, reminiscent of the "punctuated

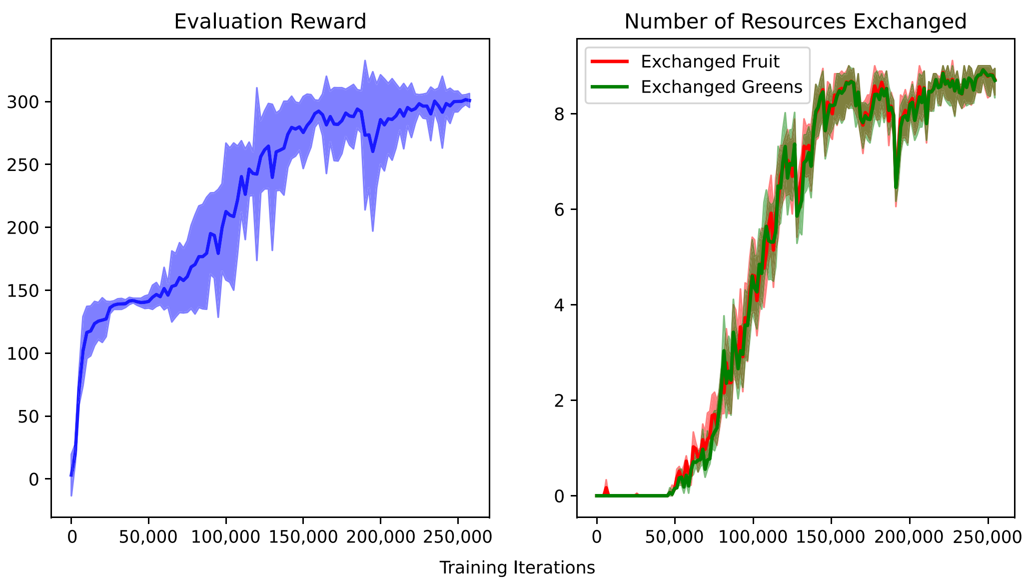

Figure 2. Cumulative reward and exchange counts over nine trials. (left) Cumulative reward over the duration of a trial, with each line representing a single trial. (right) Exchange counts averaged over nine trials; standard deviations are shown. The first steps before agents learn to avoid a large night time penalty are omitted for readability in all reward plots. We ran 10 trials total but focus on nine here, as we elect to analyze one trial separately in section 4.1.

equilibria" model of evolution (Gould & Eldredge, 1977). The first equilibrium is reached when agents learn to forage resources during the day and return to the campfire at night, reached at approximately 10,000 iterations of training and persisting for approximately 45,000 more iterations. During this period, agents do not exchange resources and instead wait out the night only, consuming just the resources they foraged and occasionally dropping a resource or moving around due to the stochastic sampling of actions during training.

AI algorithms tend to be employed on games that typically do not contain long periods of time devoted to doing nothing. Indeed, a game in which half the time is spent doing nothing would likely not be very interesting for many; however, when agents are not given an easy way to further optimize an objective, they can find novel, sometimes unexpected methods to do so, given time (Baker et al., 2020).

We observe the rise of such novel methods during the first transition approximately 55,000 training iterations in, when agents discover the ability to exchange resources around the fire. Exchange starts off in very small quantities, with agents dropping just half a unit of food in total over three nights, despite possessing an abundance of their respective resources each night. It takes thousands more iterations before the number of exchanges over three nights stabilizes around nine fruits and nine greens per episode between four agents. By this time, agents split up into two pairs, which trade with each other, as seen in Figure 3. This averages out to 3 resources exchanged per night: 1.5 resources per pair of agents. We denote this trading configuration as 2-PAIR. Each agent goes to one resource patch per day, and each patch contains five units, which all get picked. In the ideal case, each pair of agents would trade 2.5 units; however, agents need to walk from the patch back to the campfire, consuming anywhere between 0.5 and 1.0 units of food in the process. This implies that 1.5 resources per night per pair of agents is fairly close to the practical optimal quantity.

An example exchange can be found in Figure 3. Agents form into pairs and stand a cell away from their partners. Each agent drops a resource before moving over to its partner's cell to collect

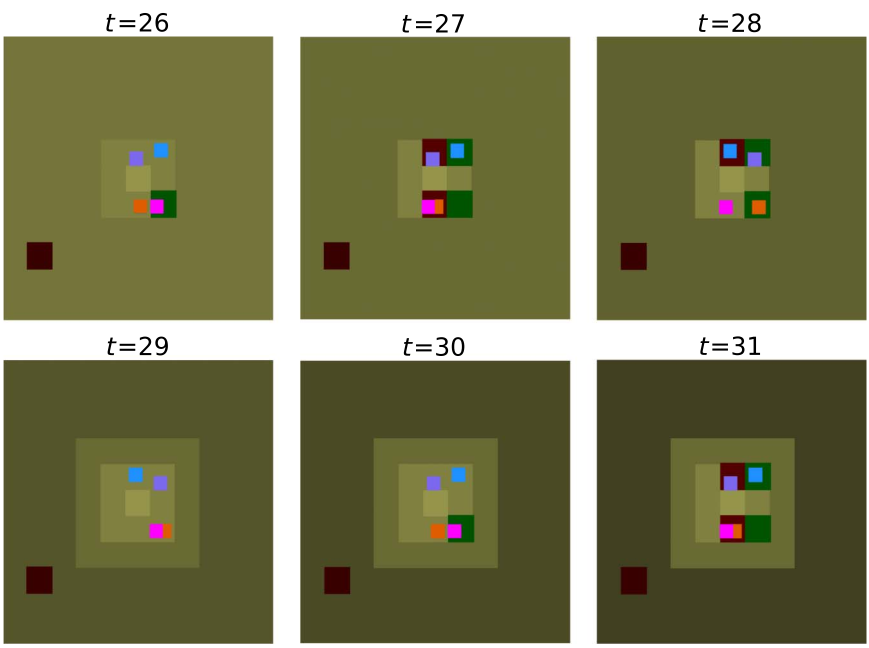

Figure 3. Five steps making up an example exchange from top left to bottom right. Four agents form into two pairs, each agent on a cell adjacent to its partner. Agents drop half a resource before moving over to the adjacent cell, where they collect the resources dropped by their partners. After collecting their partners' resources, agents move back to their original places and begin to perform another exchange.

the resource dropped by its partner. Notably, the pairs are not adjacent to each other, which may limit the degree to which different pairs can interfere with each other.

4.1 Emergence of Tolerated Theft

Across nine trials, agents will sort themselves into pairs to exchange resources, as reported. In a 10th trial however, a different form of behavior emerges that is so interesting that it deserves its own analysis.

The 2-PAIR trading configuration emerges when each agent finds its own food patch. It is possible, however, that agents do not divide themselves evenly across all patches and instead get caught in a local minimum, where two agents visit the same patch and split the resources. In this particular trial, the purple agent collects resources from both a fruit and a greens patch (accepting a minor light penalty in the evening to do so), the blue and orange agents share the other fruit patch, and the pink agent forages the other greens patch alone. As a result, the purple has significantly less incentive to trade, leaving the other three agents to divide foraged resources among each other.

The resulting behavior is fascinating; as seen in Figure 4, the pink agent will drop some of its excess greens, which lures the orange agent away from the light blue agent. The orange agent then leaves the light blue and pink agents alone to trade and collects their offering of greens. This behavior is present on every evaluation run on the final checkpoint of this trial.

To test whether this behavior is a coincidence or intentional, we take control of the pink agent to prevent it from dropping the bait and observe the change in behavior of the orange agent, which can be seen in Figure 5. The orange agent responds by interfering with the light blue and pink agents when they attempt to trade, akin to a defender in basketball. We control the pink agent during the attempted exchanges as well, because the pink agent will attempt to trade with light blue only after it drops an offering for orange. Interestingly, there is no need for the pink agent to wait for a return offer after it has left a resource to orange. This allows the orange agent to collect its resource after the pink agent has moved three cells away to trade with light blue, enabling food sharing over a distance of a few cells.

Out of all the theories for the emergence of food sharing, this behavior is the closest to tolerated theft, where agents freely give resources because the cost of defending those resources is greater than the cost of simply giving them away (Isaac, 1978). In this case, the cost of defending a resource is

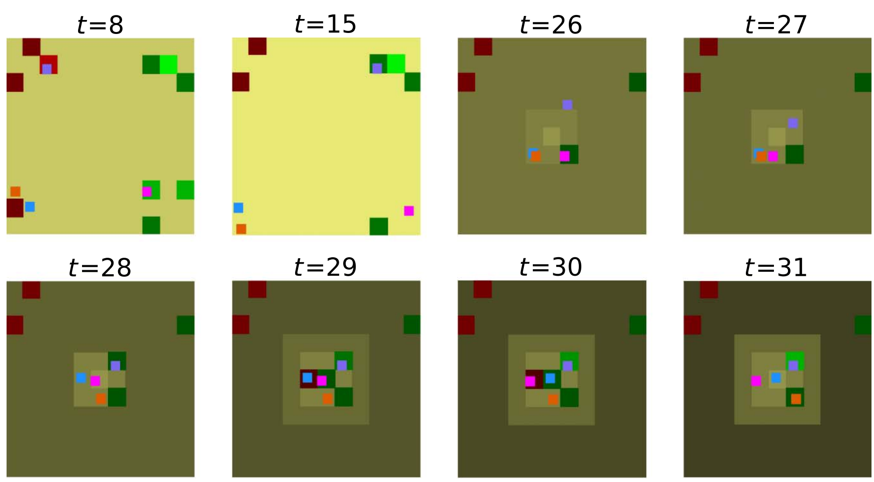

Figure 4. Key points during a concession, from top left to bottom right. For $t = 8$–15, the light blue and orange agents forage the fruits in the bottom left, the purple agent goes after the resources at the top of the map, and the pink agent gets the entire patch of greens at the bottom right. For $t = 26$–27, the pink agent drops an offering and moves toward light blue. For $t = 28$–29, orange moves toward the offering, allowing light blue and pink to begin an exchange. Purple places down greens, creating a "stockpile," as further discussed in section 4.1.1. For $t = 30$–31, orange moves to collect the offering; light blue and pink finish their exchange.

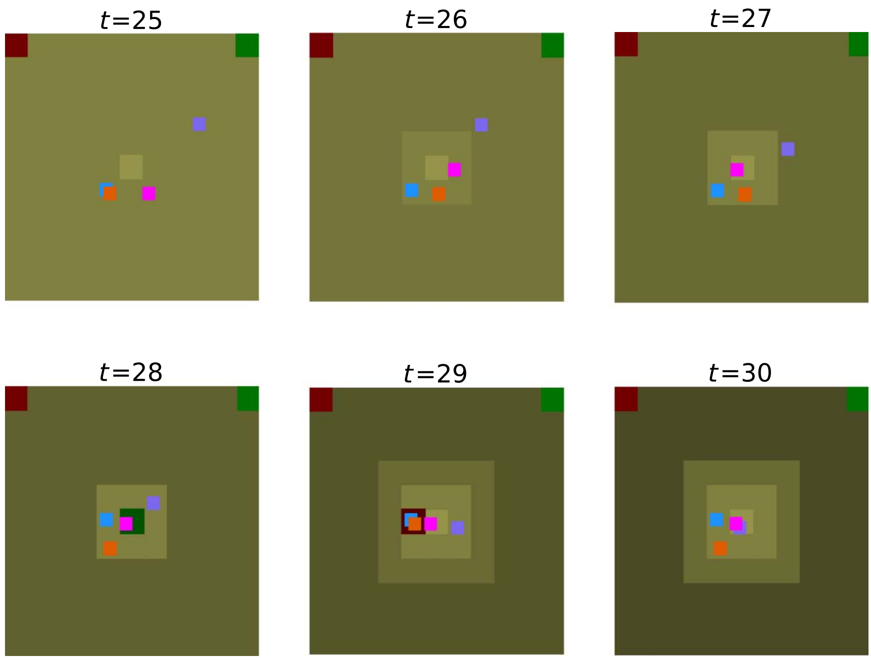

Figure 5. Orange interferes with exchange between pink and light blue agents when a concession is not made, from top left to bottom right. For $t = 25$–26, the orange agent occupies a potential exchange space. For $t = 27$–28, pink (controlled by a human) goes around the orange agent and begins to initiate an exchange with light blue. For $t = 29$, the light blue agent offers a fruit to pink, but orange moves to occupy the cell, which allows it to collect the fruit on the next turn, before pink can reach it. In response, we rescind the offer of greens. For $t = 30$, orange moves away from light blue and pink, potentially to tempt light blue and pink into another exchange.

replaced with the missed opportunity to exchange. The Trading Game supports no preprogrammed method of punishment like the punish beam described by Vinitsky et al. (2022), yet it appears that the emergence of exchange brings forth both new forms of conflict (interfering with exchanges) and new ways to deal with troublemakers (conceding resources).

4.1.1 Stockpiling Behavior

Another interesting behavior observed during the tolerated theft trial is a stockpiling behavior exhibited by the purple agent, as seen in Figure 4. In this trial, on rare occasions, the purple agent places all of its greens in its cell during the night to avoid automatically consuming them. Purple then stays on the stockpile until morning, which prevents other agents from snagging the greens in purple's absence. The next morning, before the map resets, purple picks up its unconsumed greens and then heads directly toward a fruit patch. This stockpiling behavior allows purple to save some greens for the next day, when it can consume fruits in tandem, enabling purple to achieve up to 12 additional steps of full nutritional reward for a given day. Despite the benefit of stockpiling resources, however, this behavior is inconsistent and occurs only in 6 of 30 evaluations sampled from this trial.

4.2 Resilience to Defection

In all of the experiments run, agents never exchange resources on the same grid cell; rather, they consistently stand at least one cell away from a partner and drop a resource, and if the partner reciprocates, the agents exchange spots. We call this behavior DROP-SWAP. Notably, DROP-SWAP is fairly complicated behavior, and we would not expect it to arise if the only goal between agents were exchange. A much simpler exchange strategy would involve agents meeting on the same grid cell, then dropping and picking without moving during the exchange. So why do partners always keep their distance for each exchange?

4.2.1 Intrapair Defection

We hypothesized that DROP-SWAP emerges as a mechanism to defend against defection, because a cooperative agent will have enough time to reclaim its offer before a defecting partner can grab it. We denote this kind of defection *intrapair defection*, because the cheating behavior takes place within the pair. To test this, we overwrote agent actions during evaluation to observe how agents respond when a partner reclaims its offer or attempts to grab an offer without reciprocating.

An agent can defect in two broad ways, because agents do not perform actions at the same time. The first defection occurs when an agent places down a resource but then picks it back up once its partner drops before moving to collect the partner's resource. The second defecting strategy occurs when the partner drops a resource and the agent attempts to grab it without dropping anything in return.

Across each of the nine trials, we overwrote each of the 36 agents to perform each type of defection during an exchange with its partner and measured whether it would defend against defection and rescind its offer. Out of these 36 agents, we managed to trick only three into giving up their resource. Thirty-three out of 36 agents defended their resources, moving back to their original cells if necessary, to rescind their offers. We attempted to defect again against each of the three tricked agents to determine any pattern in their behavior, but each agent defended its resource and rescinded its offer on all subsequent defection attempts. Overall, all 36 agents exhibited strong defense against defection, with only a rare slip-up every now and then. Despite our best efforts, we were unable to reliably trick any partner into giving up its food for free, supporting the hypothesis that the DROP-SWAP form of exchange arises to prevent agents from getting cheated out of their offered resources.

The ubiquity of defense against defection is rather surprising, considering the relative payoffs of getting cheated versus performing an exchange. If an agent is cheated out of half a resource without receiving anything in return, it only loses 0.5 units of total reward. Furthermore, at the time step of the defection, the loss is heavily discounted, as it only impacts future reward when agents run out of food five steps sooner. In contrast, performing an exchange yields approximately an additional 4.5 units of total reward (5 from the exchange -0.5 from running out of a resource sooner). Reward from exchange is also discounted far less, as agents immediately start receiving the reward for consuming two resources at once. Given these relative payoffs, we might not expect agents to learn to so strongly defend their resources unless it heavily motivates a partner to provide an offer in response.

4.2.2 Interpair Defection

Additionally, we noticed that across all trials, pairs never exchange adjacent to other pairs. Instead, pairs consistently trade with at least one empty cell in between them. We hypothesized that this might have emerged as a method to defend against *interpair defection*, where a pair will refuse to exchange if an outside agent is able to grab the dropped resources. To test this, we made each agent interfere with the opposite pair and measured whether it was possible to intercept resources during the exchange.

We managed to intercept at least one resource between 14 pairs out of 18 across nine trials. This does not imply, however, that stealing a resource was always easy. We observed various levels of defense: Some pairs would completely ignore outsiders, whereas others would refuse to initiate an exchange if there was an agent adjacent to the pair. In some cases, we could steal a resource from a pair right when they began their first exchange, but after getting cheated, they would refuse to exchange any further. Notably, a pair might display more defense against one outsider but completely ignore a different outsider and allow it to steal during the exchange.

We speculate that this variability in interpair defense may be a result of the difficulty in discovering exchange. If two pairs discover exchange at roughly the same time, attempting to intercept

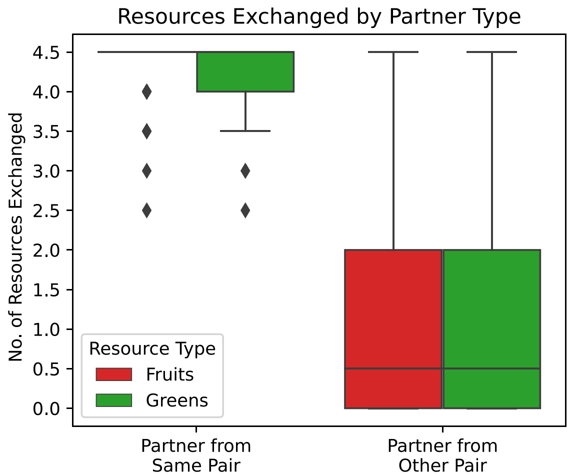

Figure 6. Number of resources exchanged over three nights between two agents. Results shown are taken from 10 evaluations for two partner pairings across nine trials, yielding 180 data points per distribution. We expect optimal agents to exchange 4.5 units of a given resource over three nights, which is approximately reached when agents trade with their preferred partners. When trading with another agent's partner, rates of exchange vary wildly.

resources from another pair will quickly prove less beneficial than exchanging with a partner. On the other hand, when one pair discovers exchange far before the other, there is heavy incentive for outsiders to get better at intercepting resources from the cooperators until the noncooperative pair can discover exchange themselves. Under this hypothesis, the difference in time between the two pairs individually discovering exchange may be an indicator of the degree of defense the first pair may have. There is no clear metric to quantify the degree to which agents defend from interpair defection, so validating this theory is not straightforward.

4.3 Interpair Cooperation

When exchanging resources in trials without tolerated theft, agents form into pairs and exchange resources with their partners. As each agent is perceived in a separate observation channel, we sought to understand the degree to which this exchange behavior is tied to a particular partner.

We measured this by freezing the normal partners from entering the campfire and seeing if agents from opposite pairs would exchange resources around the campfire if their usual partners were not available. The exchange counts over three nights can be found in Figure 6.

As in the defection experiment, the results were varied: Many interpair pairings exchanged no resources, some pairings exchanged the near-optimal amount, and other pairings would only exchange half a resource. These results imply that, despite extensive periods of time to explore interactions with other agents around the campfire, the degree to which agents explore the full range of interactions with others varies greatly.

4.4 Different Quantities

In Figure 2, we see that there is approximately a 1:1 exchange ratio between fruits and greens, which logically follows from the 1:1 distribution of resources. This prompts the following question: How do the rates of exchange change in settings with asymmetrical distributions of resources?

We ran two groups of 10 trials to answer this question. For the first group, we set the fruit and greens patches to contain six fruits and four greens, respectively; the second group had four fruits in their fruit patches and six greens in their greens patches. To focus our analysis on the exchange rate, we only analyze trials in which each agent forages a different resource patch, so a 3:2 exchange rate is always possible during the night. The 6 Fruit:4 Greens group had 8 out of 10 trials converge

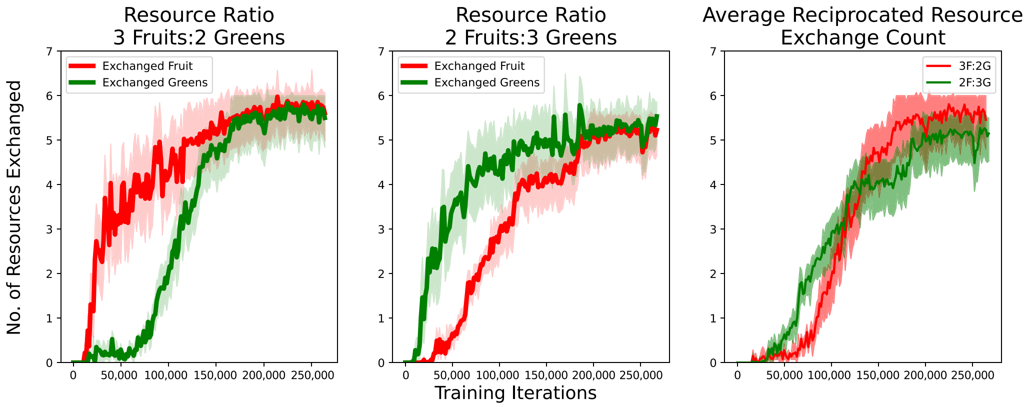

Figure 7. Exchange quantities over three nights on maps with different quantities of resources, over seven trials. Standard deviations are shown. (left) Fruit patches in this map contain six fruits, while the patches of greens contain four greens. (center) Patches of greens contain six greens, while fruit patches contain four fruits. (right) Average number of resources reciprocated, for example if agent *a* gives four fruits to *b*, but *b* only gives two greens to *a*, then two resources were reciprocated. Where the abundant resource is initially given for nothing in return, the scarcer resource is exchanged only as a form of reciprocation.

to this behavior, whereas the 4 Fruit:6 Greens group only had 7, so we conduct this analysis over 7 trials from each group.

The number of exchanges for each resource can be found in Figure 7, where we can observe interesting dynamics play out over the course of training. Initially, agents are willing to give great amounts of the abundant resource, often for nothing in return. In this setting, there is enough of the abundant resource for an agent never to run out. When an agent has so much food that it will never go a step without at least one resource to consume, dropping some of the extra resources yields the same reward as hoarding resources that will never be eaten. Thus agents with excess will occasionally drop their food, because there is no reason not to do so. To get agents to drop more than a spare resource here and there, some extra reward for doing so is required. Agents with the scarcer resource begin to offer food, which provides that reward, and so DROP-SWAP emerges. Eventually, the exchange rate approximates 1:1, with the abundant resource exchanged in slightly greater quantities than the scarcer resource.

The approach toward a 1:1 ratio is surprising. Given the 3:2 distribution of the resources, we might intuitively expect a 3:2 exchange rate to stabilize, as was the case for Johanson et al. (2022), but agents appear to stabilize on the DROP-SWAP strategy in a 1:1 ratio, despite initially giving resources away for free and possessing the ability to perform a 3:2 exchange in a single transaction. Nevertheless, after all 1:1 exchanges take place for a night, agents with the abundant resource consume whatever leftovers they have, and the cycle repeats the next day. For now, we do not attempt to provide an explanation for this phenomenon and instead leave in-depth study of this dynamic to future work.

4.5 Lowered Light Penalty

In the Trading Game, the night penalty p serves multiple purposes: (a) to pressure agents into congregating for extended periods of time and (b) to prevent agents from foraging from both sides of the map in the same day. With the rather high value $p = 10$, we've seen that agents will forage a single resource, then exchange for the other food type around the campfire; with a lower p, we might expect agents to stay out during parts of the night to forage both resources for themselves, as the weaker night penalty no longer outweighs the extra benefits from staying out to collect the other resource. Given this, we can view p as a parameter that controls the duration and degree to which agents will congregate around the fire.

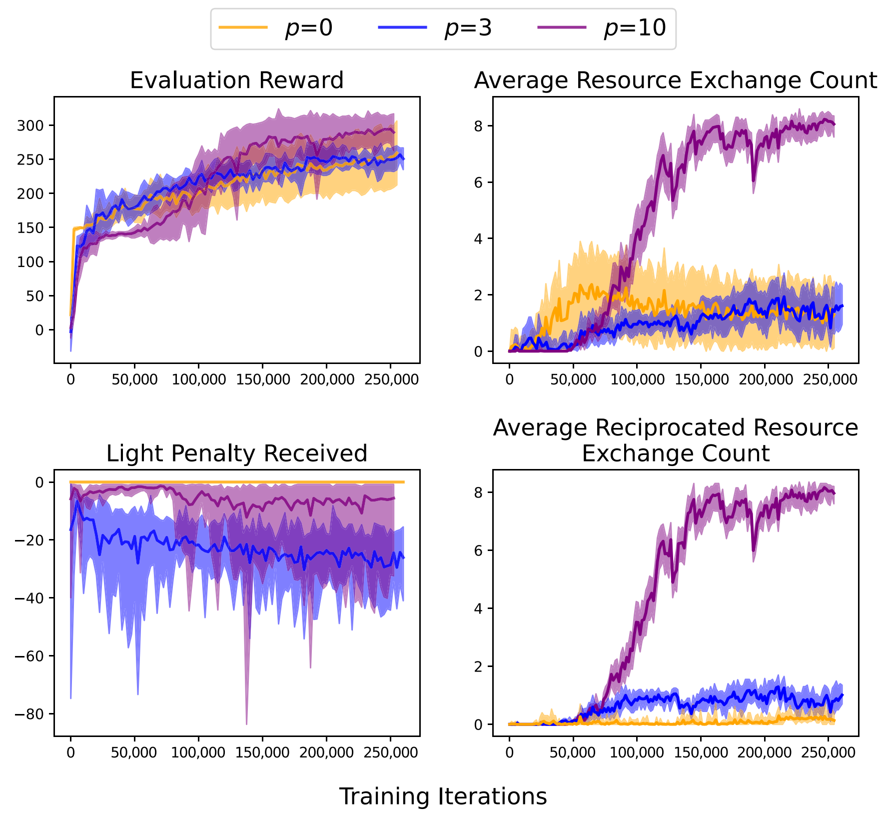

Figure 8. Metrics for $p = 0$, $p = 3$, and $p = 10$ over 10 trials each, including tolerated thefts. Standard deviations are shown. (top left) Total evaluation reward. Notably, agents that receive a larger light punishment achieve higher collective rewards than groups of agents that receive a weak light penalty and groups of agents that receive no penalty. (top right) Exchange counts with different light penalties over 10 trials; standard deviations are shown. We observe that with a lower light penalty, agents exchange less, as there is more incentive for an agent to stay out at night to collect both kinds of resources on its own. Counterintuitively, more exchanges occur when there is no light penalty than when there is a weak one. (bottom left) When $p = 3$, agents stay out during the night longer than when $p = 10$ and thus receive a greater total punishment over time. (bottom right) Number of reciprocated resources between two agents. Where we see a modest amount of exchange occur when $p = 0$, we see that very little of it is reciprocated.

To study the impact of the light penalty, we run two sets of 10 trials, one in which $p = 0$ and the other in which $p = 3$, and compare them to the 10 $p = 10$ trials. The reward and exchange counts for these trials can be found in Figure 8. We notice that agents in the lower-penalty setting exchange fewer resources, receive less total reward, and accept a larger cumulative light penalty than agents from the $p = 10$ run. Despite possessing the capability for resource exchange, agents in the $p = 3$ setting converge to the aforementioned minimum and seem to be unable to escape this minimum once it is reached. This emphasizes just how unlikely it is that resource exchange will dominate as a strategy in the Trading Game: not only must exchange be more rewarding than foraging both resources alone but also foraging both resources needs to be more costly than foraging only one.

When there is no light penalty and thus no reason to gather around the campfire, we observe very different dynamics than in the low-penalty case. A majority of the exchanges seen here are a result of agents dropping excess resources that they will never consume, as described in section 4.4. Although there are no asymmetries in the quantities of resources, it is possible for agents to forage both patches of a resource and collect more of one food type than they will ever consume. These agents will occasionally drop their excess, as there is no benefit to hoarding, allowing a one-way exchange to occur. This is made clear by the lack of reciprocation occurring during these trials.

4.5.1 Optimality of Trade

When we remove the light penalty, we alter not only the way agents explore the strategy space but the reward structure as well. We would then like to know, Does trade emerge because the light penalty makes trading optimal or because the campfire makes it far easier to discover?

Short, preliminary experiments for this work did not incorporate a campfire or day/night cycle and demonstrated no signs of emergent exchange; instead, agents from these experiments would learn to forage both resources on their own, achieving less reward than if they had traded. We hypothesized that the same outcome would emerge for $p = 0$ experiments that ran for the full 10 days.

In the $p = 0$ setting, some agents discover a strategy that provides higher reward than trading: An agent can forage an entire fruit patch, then go to an already-foraged greens patch, then place down all its fruit to create a stockpile and wait. When the greens spawn the next day, the agent can re-collect its fruit and begin immediately foraging greens. This strategy allows agents to achieve up to 10 additional steps of full-nutrition reward compared to trading, but these rewards are delayed into the following day, whereas trading can occur during the night.

Even though this stockpiling strategy is better than trading when $p = 0$, we still observe that the average reward achieved when $p = 0$ is substantially less than the average reward achieved when $p = 10$. We suspect that this is because the stockpiling strategy is difficult to learn: Whereas all $p = 10$ agents (except the ones from the tolerated theft trial) learn to trade after around 4 real-world days, only 18 out of 40 $p = 0$ agents discover the stockpiling strategy after 10 days. The other 22 agents that do not adopt the stockpiling strategy appear to compete over the remaining patches of food and rarely trade.

Thus we conclude that trade emerges due to both the changed reward structure *and* the altered exploration space imposed by the campfire. Although trade is not the optimal strategy when $p = 0$, agents will still employ strategies that perform worse than trading and seem unable to discover trading without the conditions provided by the campfire.

5 Discussion

This work takes inspiration from the concepts of autocurricula from reinforcement learning (Leibo, Hughes et al., 2019) and coevolutionary arms races (Ficici & Pollack, 1998), where agents create problems for others to solve, which then leads to the creation of clever solutions and even harder problems. In our domain, these dynamics produce the emergence of exchange; the emergence of tolerated theft-like behavior; and competitive–cooperative dynamics, such as defending exchange offers from defectors. Notably, unlike all exchange systems from previous work, the complexity of this environment arises not from agents learning to use complex game mechanics with many actions but rather from agents learning complex ways to use only nine. Given the actions of picking up and placing down resources, agents exhibit complex forms of cooperation and competition in ways we did not intend or expect. If the environment had some human-designed trading system that allowed no room for interference or defection, these dynamics likely would have not arisen.

The campfire does not explicitly facilitate exchange but instead acts as a stepping-stone (Stanley & Lehman, 2015) for its emergence by shaping the conditions under which interactions occur. In the work presented, it takes up to 4 days of wall clock time for agents to begin to reliably exchange resources on a domain where approximately 50% of the training steps are spent in a 3×3 area with four other agents. One need not imagine how long it might take for these behaviors to emerge in the $p = 0$ setting, where the chances of interacting with another agent are significantly lowered—10 days of real-world training time proved insufficient. Concepts like the "punish beam" of Vinitsky et al. (2022) or the broadcast radius of trade offers (Johanson et al., 2022) can also increase the chance of interacting with another agent, mitigating this effect, but the campfire is unique in that it emphasizes repeated interactions with the same partner during training under similar conditions. Furthermore, the campfire does not require all actions to have a large range or area of effect, allowing exchange to emerge without requiring drop and pickup actions to apply over a distance, as seen with the emergence of tolerated theft.

While the Trading Game leverages a heavy light penalty and distant resource piles to prevent agents from initially foraging both resources, the main purpose of the campfire is to promote periods of extended, idle congregation. We can conceive of a plethora of environmental modifications that could separately prevent the foraging of both resources to disentangle these two incentives, such as adding difficult terrain between resources for which agents incur a penalty to cross or making agents proficient at foraging different resource types. The goal of this work, then, is to present the Trading Game not as some benchmark task that should be accepted as-is but instead as a set of ideas that can be incorporated into and modified in other environments to study new forms of cooperation.

5.1 Informal Relationship to the Iterated Prisoner's Dilemma

Informally, the emergence of resource exchange during the night can be reasoned about in a similar fashion to the emergence of cooperation in the IPD. The night reduces the practical dimensionality of the environment, pushing agents to the small 3×3 campfire to perform whatever actions they please for the duration of the night. This setting mirrors the IPD, in which agents play multiple rounds with the same player; if agents could roam around as they pleased during the entire episode, random interactions between agents would be rare and unlikely to be iterated if agents move apart after the interaction. This may explain why resource exchange did not occur in the Fruit Market environment: A form of cooperation like modern-day markets may not need repeated interactions with the same partner, but for a behavior like resource exchange to arise, repeated interactions with previously seen individuals might be a necessary stepping-stone (Henrich, 2021; Stanley & Lehman, 2015).

We expect selfish agents first to learn to avoid dropping resources around the fire and to pick up any resources that others drop. This simple, short-term method of maximizing reward is analogous to defection in the Prisoner's Dilemma. This is indeed the first equilibrium we observe, and it persists for many thousands of training iterations.

Owing to the stochastic sampling of actions from our policies during training, agents still occasionally drop their resources around the fire, enabling agents to accidentally gift each other resources for extra reward. Agents that drop resources without receiving anything in return will avoid dropping resources in the future, which implies that if agent a wants agent b to drop a resource more often, it needs to make dropping a resource provide a higher reward for b than not dropping a resource, which it can do by offering or not offering a resource in return. This can be thought of as analogous to tit for tat, where defection and cooperation are both reciprocated such that cooperative strategies receive a higher reward on average than defecting strategies.

Like in the error-prone IPD described by Lindgren (1992), no clever agents intentionally shape other agents' behavior; rather, new forms of cooperation (like exchange) lead to new forms of defection (cheating/interfering), which leads to more complex forms of cooperation (tolerated theft/defending offers). Defection is not a viable long-term strategy because agents can alter their policies to respond to defection, so when this defection-only minimum is escaped, it must be from a cooperative strategy that is resistant to exploitation. This is likely why we always observe the rise of DROP-SWAP trading across all trials that converge to 2-PAIR.

Unlike tit for tat, however, DROP-SWAP enables cooperators to completely prevent unilateral defection, whereas tit for tat simply punishes defection on the next round. This is a key distinction between the two strategies and may be a result of the differences between the two domains. It is known that if the IPD is played a known finite number of rounds, the optimal move on the last round is to defect, which implies that the optimal move on the second-to-last round is to defect, and so on until the first round. In the Trading Game, resources are finite and episodes and nights are of a finite, fixed length. This may limit the viability of tit for tat like strategies in practice and require the guaranteed cooperation of DROP-SWAP.

5.2 Limitations

As seen in section 4.3, though agents can reliably exchange with their partners of choice, the degree to which this behavior generalizes to other agents varies. Furthermore, unlike in Fruit Market and the AI Economist, exchange in the Trading Game emerges when agents are unable to acquire both resources on their own without getting heavily penalized and are required to work together to maximize reward. This is not uncommon, however, as reference games that require cooperation to solve are used to study the emergence of natural language (Lazaridou et al., 2017).

This domain is also sensitive to the quantity of resources. If there are too few fruits, agents will consume them before they get a chance to trade; with too many fruits, agents can forage fruits one day and greens the following day, as there will be leftovers from the night before, thus reducing the benefit of mutual exchange. Furthermore, resources need to be spaced out: If fruits and greens spawned next to each other, there would be far less reason to specialize and exchange. These pitfalls could be alleviated by adding a limit to the number of resources an agent can carry at once or by adding an incentive for agents to specialize in a particular resource, as done by Johanson et al. (2022), but for the purposes of this work, we keep the environment as simple as possible.

The compute required to reach exchange emergence is also nontrivial, requiring up to 4 days of wall clock time on a Titan X[2] before a pair may begin to reliably trade and up to 10 days in experimental settings with eight agents. This bottlenecked iteration speed and made it difficult to predict when an experimental configuration could yield emergent exchange or not, as performance does not improve significantly during the first equilibrium. Environments supporting greater social complexity with larger numbers of agents would likely take significantly longer.

5.3 Future Work

Given the results and limitations described in the Trading Game, there remains ample room for future work. Within the Trading Game, many interesting environmental properties remain to study, such as the potential effects of adding a carrying capacity for resources or relative food quantities. Scaling to larger numbers of agents and generalizing this exchange behavior to all seen and even unseen agents are also of great interest.

The simplicity of the Trading Game also makes it ripe for extension; for example, the addition of a local communication system could allow agents to negotiate around the campfire. Deep neuroevolutionary approaches have been successfully applied as competitive alternatives to single- and multiagent reinforcement learning problems (Klijn & Eiben, 2021; Such et al., 2018) and show potential as another algorithm for the study of emergent cooperation. Concepts analogous to the campfire could be applied to related domains, such as Fruit Market, and even to entirely different domains, such as reference games for research on emergent communication (Lazaridou & Baroni, 2020).

6 Conclusion

In this work, we demonstrated how novel behaviors can arise by reshaping the environmental conditions under which agents interact. We showed that a simple foraging environment with periodic gathering around a campfire can lead to emergent trading behavior and discussed how the emergence of trading in our setting is analogous to the evolution of cooperation in the IPD. By directly interacting with the agents, we found that agents could prevent themselves from being cheated by their usual partners, but they exhibited varying levels of defense against being cheated by a third party. Additionally, we observed that agents could learn to interfere with exchanges as an indirect form of punishment, allowing an emergent behavior similar to tolerated theft to emerge. As

2 Although an older GPU, the Titan X has a theoretical performance of 6.69 TFLOPS and is no slacker by any means. Test experiments run on a V100 saw only an approximate 25% iteration speedup.

congregation pressure is reduced, these behaviors arise in much weaker forms, if it all, demonstrating the importance of extended congregation to the emergence of embodied cooperation.

Acknowledgments

J.G. has been supported by a grant from the Fourmentin Foundation. We also acknowledge computational support from the Brandeis HPCC, which is partially supported by the NSF through DMRMRSEC 2011846 and OAC-1920147. We thank Thomas Willkens for many productive conversations and comments throughout this project, as well as Michael Johanson and Joel Liebo for comments on early versions of this work. We also thank the anonymous reviewers for their insightful feedback, which improved the article.

References

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. J., Johanson, M. B., Singh, S., Haas, J., Mordatch, I., Mobbs, D., & Leibo, J. Z. (2023). *Melting Pot 2.0*. ArXiv. https://doi.org/10.48550/arXiv.2211.13746

Axelrod, R. M. (1984). *The evolution of cooperation*. Basic books.

Axelrod, R. (1987). The evolution of strategies in the iterated prisoner's dilemma. In L. Davis (Ed.), *Genetic algorithms and simulated annealing* (pp. 32–41). Morgan Kaufman.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2020). *Emergent tool use from multi-agent autocurricula*. ArXiv. https://doi.org/10.48550/arXiv.1909.07528

Bhoopchand, A., Brownfield, B., Collister, A., Lago, A. D., Edwards, A., Everett, R., Frechette, A., Oliveira, Y. G., Hughes, E., Mathewson, K. W., Mendolicchio, P., Pawar, J., Pislar, M., Platonov, A., Senter, E., Singh, S., Zacherl, A., & Zhang, L. M. (2022). *Learning robust real-time cultural transmission without human data*. ArXiv. https://doi.org/10.48550/arXiv.2203.00715

Burtsev, M., & Turchin, P. (2006). Evolution of cooperative strategies from first principles. *Nature*, *440*(7087), 1041–1044. https://doi.org/10.1038/nature04470, PubMed: 16625195

de Witt, C. S., Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P. H. S., Sun, M., & Whiteson, S. (2020). *Is independent learning all you need in the StarCraft multi-agent challenge?* ArXiv. https://doi.org/10.48550/arXiv.2011.09533

Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Brookings Institution Press. https://doi.org/10.7551/mitpress/3374.001.0001

Ficici, S. G., & Pollack, J. B. (1998). Challenges in coevolutionary learning: Arms-race dynamics, open-endedness, and mediocre stable states. In *Proceedings of the sixth international Conference on Artificial Life*. (pp. 238–247). Association for Computing Machinery.

García, J., & van Veelen, M. (2018). No strategy can win in the repeated prisoner's dilemma: Linking game theory and computer simulations. *Frontiers in Robotics and AI*, *5*, 102. https://doi.org/10.3389/frobt.2018.00102, PubMed: 33500981

Gauthier, J., & Mordatch, I. (2016). *A paradigm for situated and goal-driven language Learning*. https://doi.org/10.48550/arXiv.1610.03585

Gostoli, U., & Silverman, E. (2023). Self-isolation and testing behaviour during the COVID-19 pandemic: An agent-based model. *Artificial Life*, *29*(1), 94–117. https://doi.org/10.1162/artl_a_00392, PubMed: 36269874

Gould, S. J., & Eldredge, N. (1977). Punctuated equilibria: The tempo and mode of evolution reconsidered. *Paleobiology*, *3*(2), 115–151. https://doi.org/10.1017/S0094837300005224

Gupta, A., Lanctot, M., & Lazaridou, A. (2021). *Dynamic population-based meta-learning for multi-agent communication with natural language* [paper presentation]. 35th Conference on neural information processing systems, online.

Henrich, J. P. (2016). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press. https://doi.org/10.1515/9781400873296

Henrich, J. P. (2021). *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Picador.

Hinsch, M., & Bijak, J. (2023). The effects of information on the formation of migration routes and the dynamics of migration. *Artificial Life*, *29*(1), 3–20. https://doi.org/10.1162/artl_a_00388, PubMed: 36383052

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Hughes, E., Leibo, J. Z., Phillips, M. G., Tuyls, K., Duéñez-Guzmán, E. A., Castañeda, A. G., Dunning, I., Zhu, T., McKee, K. R., Koster, R., Roff, H., & Graepel, T. (2018). *Inequity aversion improves cooperation in intertemporal social dilemmas*. ArXiv. https://doi.org/10.48550/arXiv.1803.08884

Isaac, G. (1978). The food-sharing behavior of protohuman hominids. *Scientific American*, *238*(4), 90–108. https://doi.org/10.1038/scientificamerican0478-90, PubMed: 418504

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P. A., Strouse, D. J., Leibo, J. Z., & de Freitas, N. (2019). *Social influence as intrinsic motivation for multi-agent deep reinforcement learning*. ArXiv. https://doi.org/10.48550/arXiv.1810.08647

Johanson, M. B., Hughes, E., Timbers, F., & Leibo, J. Z. (2022). *Emergent bartering behaviour in multi-agent reinforcement learning*. ArXiv. https://doi.org/10.48550/arXiv.2205.06760

Kaplan, H., Hill, K., Cadeliña, R. V., Hayden, B., Hyndman, D. C., Preston, R. J., Smith, E. A., Stuart, D. E., & Yesner, D. R. (1985). Food sharing among ache foragers: Tests of explanatory hypotheses [and comments and reply]. *Current Anthropology*, *26*(2), 223–246. https://doi.org/10.1086/203251

Klijn, D., & Eiben, A. E. (2021). *A coevolutionary approach to deep multi-agent reinforcement learning*. ArXiv. https://doi.org/10.48550/arXiv.2104.05610, https://doi.org/10.1145/3449726.3459576

Lazaridou, A., & Baroni, M. (2020). *Emergent multi-agent communication in the deep learning era*. ArXiv. https://doi.org/10.48550/arXiv.2006.02419

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). *Multi-agent cooperation and the emergence of (natural) language*. ArXiv. https://doi.org/10.48550/arXiv.1612.07182

Leibo, J. Z., Hughes, E., Lanctot, M., & Graepel, T. (2019). *Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research*. ArXiv. https://doi.org/10.48550/arXiv.1903.00742

Leibo, J. Z., Perolat, J., Hughes, E., Wheelwright, S., Marblestone, A. H., Duéñez-Guzmán, E., Sunehag, P., Dunning, I., & Graepel, T. (2019). *Malthusian reinforcement learning*. ArXiv. https://doi.org/10.48550/arXiv.1812.07019

Lindgren, K. (1992). Evolutionary phenomena in simple dynamics. In C. G. Langton, C. Taylor, J. D. Farmer, & S. Rasmussen (Eds.), *Artificial Life II* (Vol. 10, pp. 295–312). Addison-Wesley.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2020). *Multi-agent actor-critic for mixed cooperative-competitive environments*. ArXiv. https://doi.org/10.48550/arXiv.1706.02275

McKee, K. R., Hughes, E., Zhu, T. O., Chadwick, M. J., Koster, R., Castaneda, A. G., Beattie, C., Graepel, T., Botvinick, M., & Leibo, J. Z. (2021). *Deep reinforcement learning models the emergent dynamics of human cooperation*. ArXiv. https://doi.org/10.48550/arXiv.2103.04982

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative agents: Interactive simulacra of human behavior*. ArXiv. https://doi.org/10.48550/arXiv.2304.03442, https://doi.org/10.1145/3586183.3606763

Park, J. S., Popowski, L., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2022). *Social simulacra: Creating populated prototypes for social computing Systems*. ArXiv. https://doi.org/10.48550/arXiv.2208.04024, https://doi.org/10.1145/3526113.3545616

Pricope, T.-V. (2021). *Deep reinforcement learning in quantitative algorithmic trading: A review*. ArXiv. https://doi.org/10.48550/arXiv.2106.00123

Rashid, T., Samvelyan, M., de Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2018). *QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning*. ArXiv. https://doi.org/10.48550/arXiv.1803.11485

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*. ArXiv. https://doi.org/10.48550/arXiv.1707.06347

Stanley, K. O., & Lehman, J. (2015). *Why greatness cannot be planned: The myth of the objective*. Springer International. https://doi.org/10.1007/978-3-319-15524-1

Suarez, J., & Isola, P. (2022, April 25–29). *Specialization and exchange in neural MMO* [Paper presentation]. In 10th international conference on learning representations, online.

Such, F. P., Madhavan, V., Conti, E., Lehman, J., Stanley, K. O., & Clune, J. (2018). *Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning*. ArXiv. https://doi.org/10.48550/arXiv.2106.09012

Vinitsky, E., Köster, R., Agapiou, J. P., Duéñez-Guzmán, E., Vezhnevets, A. S., & Leibo, J. Z. (2022). *A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings*. ArXiv. https://doi.org/10.48550/arXiv.2106.09012, https://doi.org/10.1177/26339137231162025

Yu, C., Velu, A., Vinitsky, E., Wang, Y., Bayen, A., & Wu, Y. (2021). *The surprising effectiveness of PPO in cooperative, multi-agent games*. ArXiv. https://doi.org/10.48550/arXiv.2103.01955

Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., & Socher, R. (2020). *The AI economist: Improving equality and productivity with AI-driven tax policies*. ArXiv. https://doi.org/10.48550/arXiv.2004.13332

Appendix

We use RLlib provided in Ray 1.13.0 and PyTorch 1.11.0 in our experiments. All parameters not mentioned in Table A1 can be assumed to be the default for these versions.

Table A1. PPO hyperparameters.

Hyperparameter	Value
Learning rate	1e-4
Gamma	0.99
GAE λ	0.95
Clip range	0.03
Entropy coefficient	0.05
Value function coefficient	0.25
Num SGD iterations	5
Batch size	2,000
Minibatch size	2,000
LSTM max sequence length	50