Melting Pot Cohort 68 : Exploring MARL Safety in meltingpot.

Authors : Gema Parreño, Peter Francis, Cam Tice, Chris Pond, Yohan Mathew, Tomasz Steifer, Marina Levay

Repository , Video

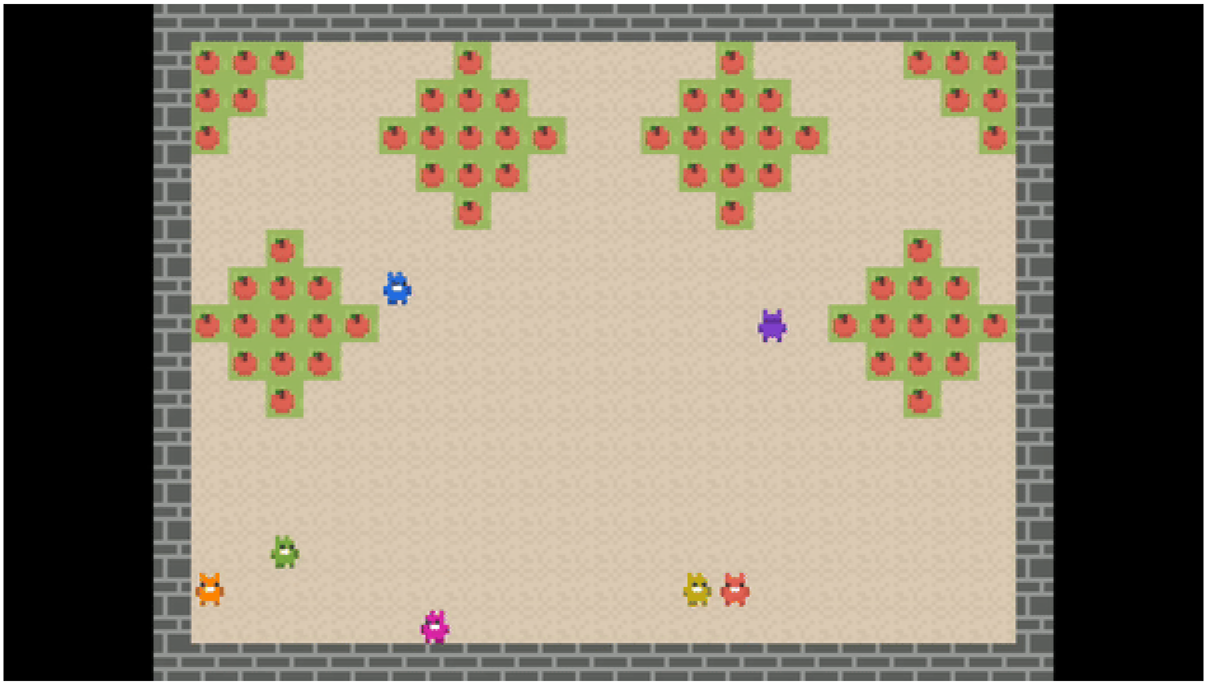
WHAT WE DID

This ongoing working document was produced as the final project of the AI Alignment Course by Blue Dot Impact and has a first due date in June 2024. The project is based on a Multi Agent Reinforcement Learning **simulation that explores the tragedy of the commons dilemma**. We explored the following research questions:

- *In scenarios where the tragedy of the commons can be assumed, what elicits cooperation in multiagent systems?*
- *What kind of relevant AI Safety insights can we extract from exploring cooperative multiagent systems?*

To discuss this question, we conducted several experiments that:

1. *Introduced evaluations*. Measuring **agent generalization and capabilities** in various setups , comparing focal population performance across scenarios, and focal population versus background population.
2. Introduced *changes to the game environment and dynamics*. By changing resource respawn rates we examined the **risks of overharvesting** and the agents' ability to maintain equilibrium in scarce and abundant environments. By modifying the environment we tested the **resilience of AI systems** to environmental changes, introducing concepts like private property.
3. *Disabled the ability of agents to punish each other*. We examined the **risks associated** with a **lack of punitive measures** and removing social norms enforcement. We called this agent *no_zap* , and created a specific substrate called commons_harvest_disabled_punishment
4. *Modified the reward signal during training*. From an AI Safety standpoint, this experiment examines the potential for **misaligned incentives** leading to selfish behavior. We called these agents Farmer, and created a specific substrate called *commons_harvest_farmer* , together with a specific farmer lua component.
5. Had discussions about **risks in Multiagent systems** and highlighted one specific situation within our exercise in melting-pot and the tragedy of the commons dilemma..



Simulation 1. Commons harvest open substrate mimicking the tragedy of the commons dilemma. Trained under a set of fully independent learning agents. In this game, agents must collect apples while ensuring the sustainability of the apple field. If the last apple disappears, the apple field is depleted. The agents can zap other agents, enabling punishment as a social norm. This is considered a mixed-motive game that balances competitive with cooperative efforts, as the agents must gather the highest number of apples (competition) but also must let the apple field regrow (cooperation)

WHY COOPERATION?

According to Dafoe et al[1], Cooperation plays a big part in humanity's progress and success, and AI that cooperates is fundamental in a world where multiagent interactions will be a reality. Some mental models point to the fact that *crucial crises confronting humanity are challenges of cooperation* [2]. Some argue that a multi-agent learning approach may be considered a form of superintelligence necessary to ensure a beneficial net outcome in automated processes [3 Collective SupperIntelligence]. In light of this, we believe that understanding cooperation can be fundamental to reducing AI-related risks.

The intersection between complex systems that require cooperation and AI has been proposed as a research agenda by several experts [4][5], gaining traction in mixed-motive games such as Diplomacy [6] and melting-pot [7]. These games present a real multidimensional complexity of both cooperative and competitive dynamics that mimics reality [8] and have been studied and classified in the AI community [9] [10] from some interesting perspectives such as reputation [11], communication [12] [13], and disagreement [14] to influence agents behavior.

The commons dilemma

The tragedy of the commons studies the tension between collective and individual rationality[15][16], or *how individual choices can affect collective loss*: it is used to describe what happens when individuals use or gather a shared resource—apples, in the case of commons harvest substrate—for their own benefit without considering the impact of their actions on the wider community. Over time, this selfish use can lead to the depletion or destruction of that resource, making it less available or even unusable for everyone. This dilemma highlights how individual interests, when not aligned with the common good, can lead to the ruin of shared resources and worse outcomes for all individuals.

HOW DOES COOPERATIVE AI RELATE TO AI SAFETY?

The majority of AI Safety work is concerned with single-agent scenarios (individual alignment). The situation is qualitatively different in the multi-agent scenarios[17] where the agents have to respond to other agents' behavior, which is hard or impossible to control, or, for instance, where agents can communicate in a way that is incomprehensible to the human observer.

As agents become more intelligent and capable of doing the tasks they are assigned to do, they are also more capable of doing harm and deceiving people. Center on Long Term Risks´ differential progress research agenda [18] argues for the need to improve multi-agent cooperative capabilities in a way that does not significantly increase harmful ones. This is important because cooperating capabilities can be detrimental to social welfare, as the underlying agents' understanding that leads to cooperation can facilitate deception and coercion.

There are also several coordination challenges for preventing AI conflict[19]: Transformative AI scenarios involving multiple systems pose a unique existential risk of catastrophic bargaining [17][20], a failure between multiple AI systems (or joint AI-human systems). It's possible that we can't delegate its solution to individual alignment, which poses interesting questions: *Is individual alignment enough? What happens in terms of safety in a multi-agent transformative AI scenario?* We are interested in the safety concerns related to cooperation, broader risks, and experiments derived from questions.

WHAT WE FOUND

The following section goes over our interpretations of the experiments done and requires some knowledge about the melting pot generalization framework: we recommend the reader go to the Evaluation Criteria of Generalization and Capabilities if they are not familiar with the Melting pot framework.

We believe these conclusions are preliminary and could be updated by doing more research, by e.g increasing the number of episodes of the evaluations or training better agents, so we take them as preliminary insights.

1. We were not able to **outperform our trained baseline**, and our agents were **not more sustainable**. What **can we do next?**

We measured the reward of all or trained agents over different substrates and under different mixes of focal and background populations. When we plot the histograms of these rewards we see that none of our agents is clearly outperforming the baseline agent (open). The **Y-axis** shows the number of occurrences (frequency) of rewards that fall within each bin of the histogram. Each bin on the **X-axis** represents a range of reward values.

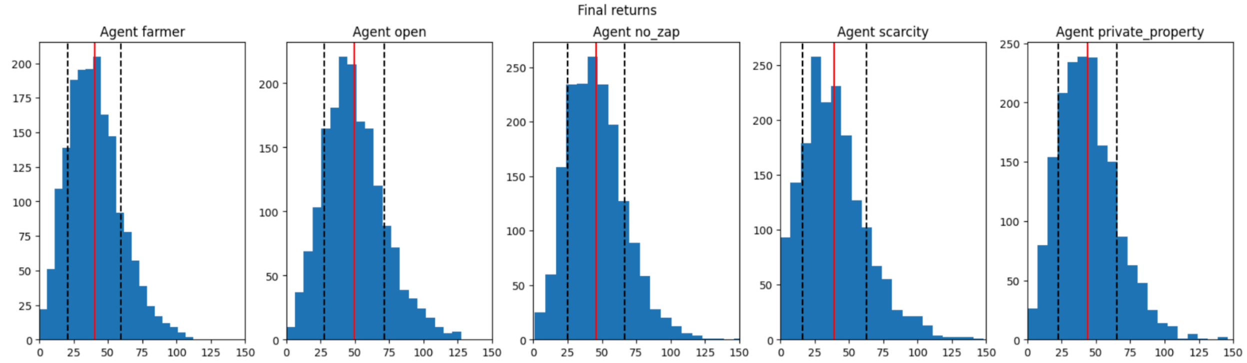


Fig1. By comparing the mean and the std of the rewards for different agents, you can get a sense of which agent tends to perform better or more consistently. For instance, if one agent has a higher mean and a smaller standard deviation compared to others, it suggests that this agent consistently achieves higher rewards.

Insights :

- All agents have similar rewards, which suggests that on **average, they perform similarly under the given scenarios.**
- The standard deviations are also similar across agents, indicating that the variability in their performance is comparable.
- The overall distribution shapes are quite similar, suggesting that the scenarios and experimental setups affect all agents in a relatively uniform manner.

The **next steps** we propose to overcome this challenge are:

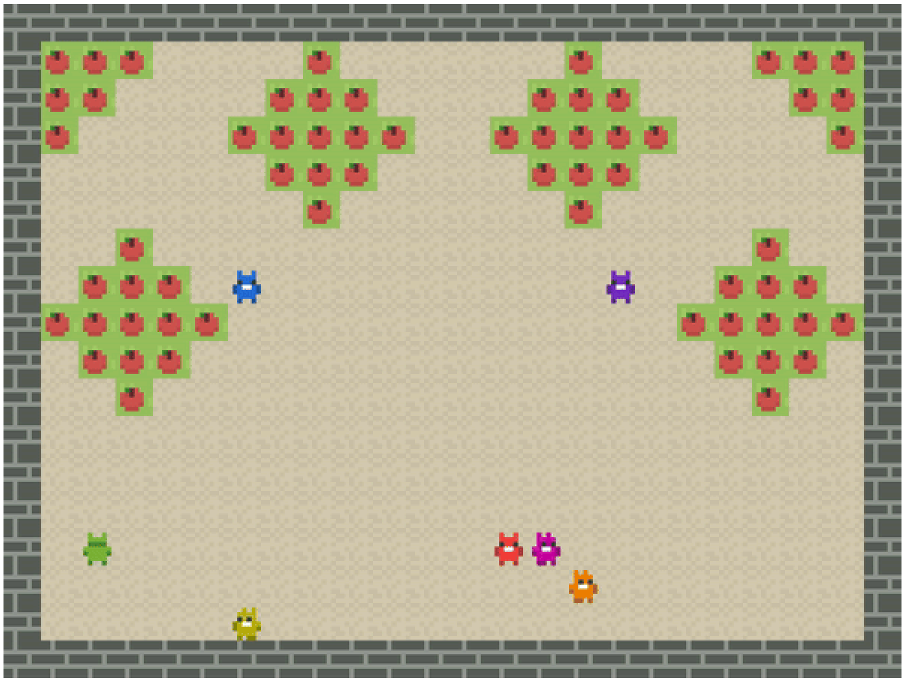
  - **Farmer**. Increase reward for observing apples. Changing the reward by an order of magnitude.
  - **No zap.** Include the reward for farmer, coexisting with disabling of the punishment mechanism. Including farmer ideas into a disablement punishment scenario.
  - **Open**. Retrain open agents with rewards designed strictly for punishment conditions, with reward for punishment and penalty for being zapped
  - **NN Architecture change**. Change the LSTM from 32 to 128 and match DeepMind's Melting Pot CNN Architecture, as this doesn´t bottleneck training
  - Use **scarcity** and **private_property** created substrates for evaluation only.

2. It is important to consider collusion in a MultiAgent Safety Scenario when punishment is disabled.

Our results point to a new direction to study agent behavior with respect to punishment as a social norm in the absence of the ability to do it during training and their response when punishment is available during evaluation. It intends to measure **robustness against punishment** when it is disabled and the study of asymmetries in punishment inside the action space in multi-agent reinforcement Learning.

Why is this important for AI Safety?

- Lack of studies about empirical asymmetrical situations in social norms in Multi-agent Reinforcement Learning.
- Need for conclusions concerning what happens when agents without punishment coexist with ones that do punish.
- Offers a testbed for testing policies that might be robust against aggressive agents.



Simulation 2. Melting pot *commons_harvest__open substrate*. The focal agent population is trained in the simulation without having punishment (zap) in their action space. An example of a Multi-agent Reinforcement Learning collusion scenario that involves restricting action spaces.

3. We can improve our evaluation methods

Even though we managed to create our own set of focal vs background populations and were able to extract insights about evaluation, we think the project could benefit from:

- **Measuring the number of apple fields depleted** could differentiate between agents that are not gathering apples but letting the apple field re-spawn, as right now some evaluation insights might lead to confusion concerning the agent response to the dilemma
- Carefully **craft the focal population scenarios** we are interested in, and run evaluation for more episodes.
- Choose a **more statistical representation** during evaluation so we can get more robust insights.
- Measuring the number of **useful zaps** during evaluation. We didn't have a metric for measuring punishment actions.

Insights trained agents

- The best performance of trained agents (focal) is in the default substrate with no background population.
- Substrates that had agents trained on open baseline or those in which punishment was disabled were the ones in which agents showed less unequal agent behavior
- **Highlighted agent**: No_zap :
    - No default environment: apple field not depleted in eval with 10 episodes during 3000 timesteps.
    - Smaller percentage of dead substrates in all environments with default regrowth rate. (behind open agent) - in between 10 % and 20%)
    - Final returns closer to open benchmark in average return (50)
    - Best focal population performance with respect to background sustainable visitors populations (agents trained by Deepmind).
    - Outperforming in all environments as a focal population with respect to farmer agents as the background population.
    - Outperforming as the focal population in 3 environments with respect to open agents as the background population.
    - Outperforming in all environments as a focal population with respect to scarcity as background agents.
- **Highlighted agent:** Farmer.
    - Best performance in a default scenario. Highest average reward and less inequality among agent behavior.
        - Worst performance Highest overall percentage in dead substrates in all environments (all above 40%)
        - Final returns are slightly behind the open benchmark in average return (50)
        - Outperforming as a focal population in all substrates with scarcity as background agents
        - Outperforming as a focal population in 4 substrates with no zapping as background agents
        - Outperforming as a focal population in all substrates with scarcity as background agents

Insights rest of the agents

- SCARCITY
    - Worst default performance. Apple field depleted after 1000 timesteps.
    - **Evaluation experiments with respect to regrowth rate**: from 100x less regrowth to the baseline level of regrowth, we see the level of reward change relatively slowly, and inequality stays relatively high. However, once we increase the rate of respawn by 10, we see almost a 5X increase in reward and a high drop in inequality.

- Outperforming as the focal population in 3 environments with farmers as background population
- Outperforming as the focal population in 4 substrates with open as background agent.

Insights framework

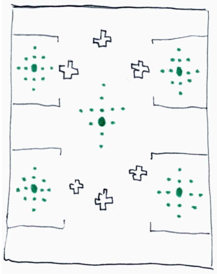
Find below some insights about the Melting-pot framework that we found useful for our exploration and that might save time to people that are currently exploring the framework.

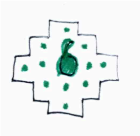
- **Create specific plots** that measure what you are interested in. (eg: the dilemma)
- Set the background population evals dynamics and scenarios from the start.
- What worked: get the dilemma into the evaluation in plots (barchats) and apple depletion in curve analysis.
- Compare the background population with the focal population of our agents.
- Compare sustainable background population with focal population trained by Deepmind. The best focal population should perform better with a sustainable background population.
- FocalVSBackground can bring interesting situations that might affect safety.
- Meltingpot framework understanding and control is non-trivial and has some learning curve.
- The background population of common harvest substrate trained from DeepMind is trained on common_harvest_closed substrate.

EXPERIMENTS OVERVIEW

Table conclusion experiments

Exp Name	Idea	Safety Impact / Discussion	Conclusions
No_zap	**Disabling zapping.** Make it so that all agents (focal and background population) cannot zap others.	Having access to the zap action allows agents to punish others and defend territories. In all Melting Pot scenarios, it is assumed that the zap action is necessary to maintain “order”,	Found some insightful scenarios regarding asymmetries between agents that zap and agents that don't. In terms of response with respect to the dilemma, it was the best agent performing 10% below our trained

		and to not lead to the extinction of the commons. This could help understand if restricting actions can lead to a more cooperative agent and explore Collusion in Multiagent systems	baseline. Future directions include retraining this agent with reward design coming from farmer agent, and studying more about its punishment conditions.
Farmer	Establish a **reward** for ensuring respawn. Establish a reward system that encourages agents to leave the last apple to regrow, fostering sustainability. 	Adjusts in-game rewards to promote sustainable harvesting and cooperation among agents. Encourages agents to conserve resources by not harvesting the last available resource, thereby allowing it to regrow and maintain resource availability for the collective.	The results were fairly inconclusive due to the low reward given during training. Future directions include re-training and improving in 1 or 2 the order of magnitude the designed reward for observing apples. Future directions include retraining baseline with change in punishment actions and comparing focal VS background populations in this setup.
Private_property	Changing the **location of apple fields** and adding walls in the Common Harvest substrate  Change game setup to find interesting situations (number of apple fields, apples/agents ratio) and wall location	This change affects how the environment itself constrains or guides agent behavior, by structuring physical barriers and resource locations which provide immediate visual and strategic feedback to the agents. Although this experiment primarily modifies physical environment structure, it indirectly relates to how resources are managed and accessed, influencing cooperative strategies by altering resource availability and accessibility	The default behavior underperformed the rest of the agents and some other substrates as a partnership and closed inside melting-pot could be more insightful. No more agents will be trained on this substrate and it will be used for evaluation only purposes.
Abundance/Scarcity	Changing the **apple respawn radius** and growth probabilities	This directly aligns with adjusting how resources respond to agent behaviors. By changing the regrowth probabilities, we alter the environmental dynamics to	This substrate has been proven interesting during evaluation to measure generalization toward response to resource changes, In the future, this substrate will be used

	 :	either encourage or discourage certain behaviors, aiming to promote sustainable resource management	for evaluation only.

Evaluation Criteria of Generalization and Capabilities

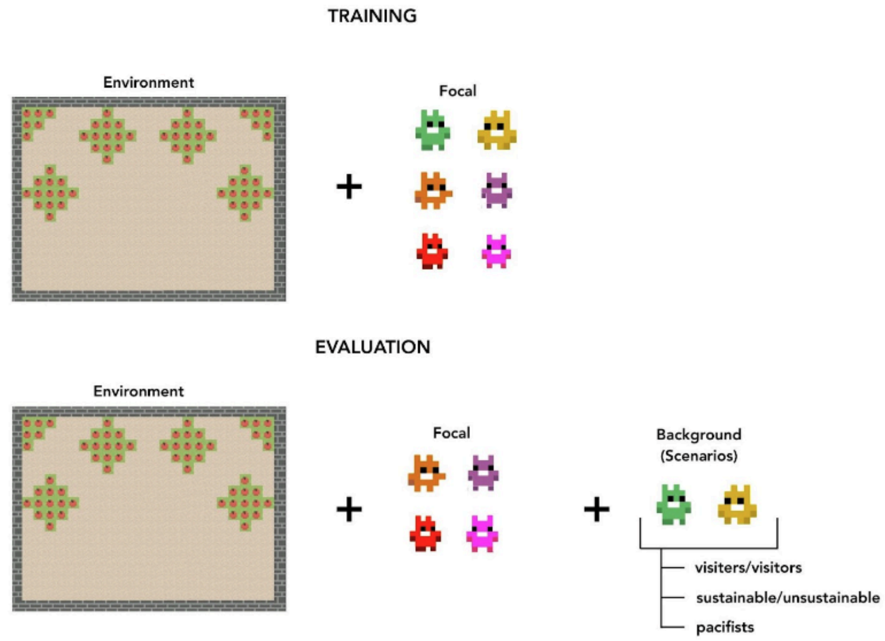


Diagram 1. Meltingpot framework evaluation generalization. During training, the focal population learns in the environment. During evaluation, a pre-trained background population with different dynamics is introduced. This allows us to compare how trained agents (focal) respond to unseen dynamics in other agents (background). We first explored sustainable/unsustainable and pacifist dynamics but all results shown in the exercise only contain visitors .

To understand evaluation inside the melting-pot framework, we note the generalization capabilities that melting-pot framework offers and highlight the difference between the **focal population** and the **background population**: in essence, the focal population is the primary group of agents whose adaptability and generalization are measured, while the background population serves as the unfamiliar social partners introduced during testing to create diverse and unpredictable social scenarios.

- **Focal population**: the focal population consists of the agents that are being evaluated for their ability to generalize to novel social situations. These agents are trained with access to the physical environment (substrate) but without any exposure to the individuals in the background population during their training phase. The performance of the focal population is measured in test scenarios to determine how well these agents can adapt to social situations involving both familiar and unfamiliar individuals
- **Background population**: The background population gathers the set of agents that the focal population encounters during the test scenarios. These background agents are designed to create new social dynamics and challenges that the focal population has not experienced during training. By mixing the focal population with the background population, the evaluation aims to test the generalization capabilities of the focal agents to novel social interactions. Among the background population, we can find 4 subgroups that create different scenarios: *visitor and visit*, in which we alter the distribution of focal and background agents, creating a dynamic in which trained agents "visit" a different population or the trained agents receive visitors. Pacifists and zappers;

In order to measure generalization capabilities, we combine the different focal populations with the environments and different scenarios, producing a set of evaluations that measure in different capabilities.

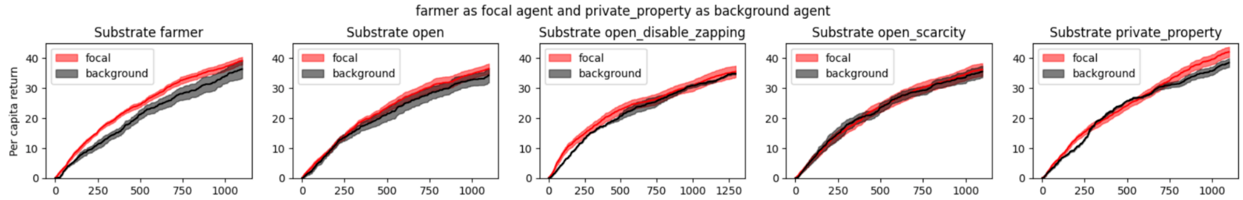
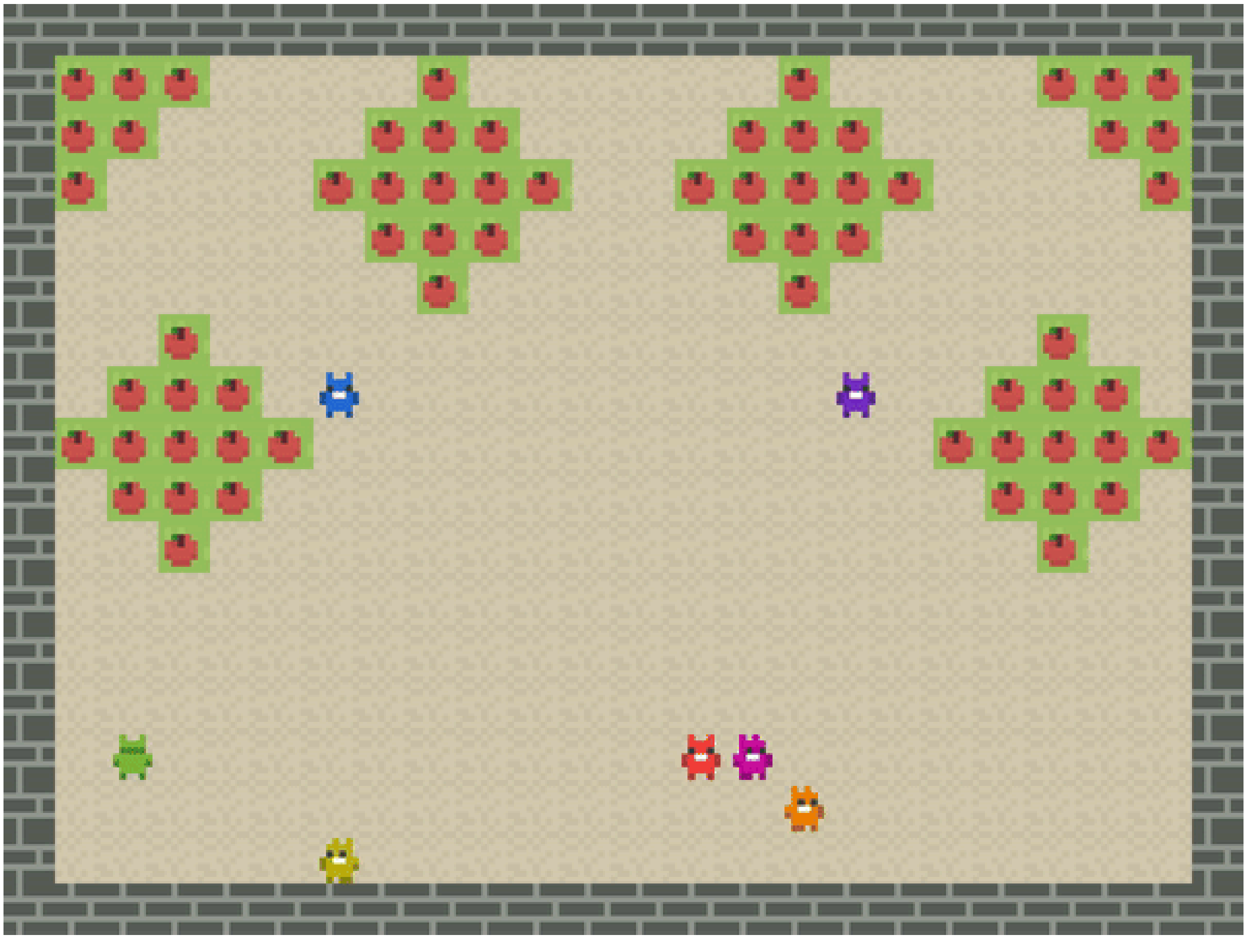


Fig2. Evaluation of focal population VS background population performance. In a nutshell, we put 5 agents trained with farmer conditions and 2 coming from private property and evaluated them in different environments.

Experiment 1 . Disabling Punishment

Disabling the punishment mechanism tests the impact of removing social norms enforcement. This experiment is significant for AI Safety as it examines **the risks associated with a lack of punitive measures**, such as the breakdown of cooperation and an increase in selfish behaviors. It helps in understanding the importance of social norms and punishment in maintaining safe and cooperative multiagent systems. This perspective is in line with the discussions on scalable supervision and the enforcement of social norms to ensure safe behavior in AI Systems by Amodei et al. (2016)[22]



Simulation 3. Evaluation Focal population of agents without zapping + background population of agents able to zap (green+yellow agents) in commons__harvest: open substrate. A set of 5 agents have been trained without the punitive social norm action, and then are evaluated together with agents that are able to punish.

Even though we highlighted this agent as interesting for potential studies, overall performance was not shown best overall. However, we encountered interesting behaviors of no_zapping agents when they faced both substrates in which punishment was enabled and disabled.

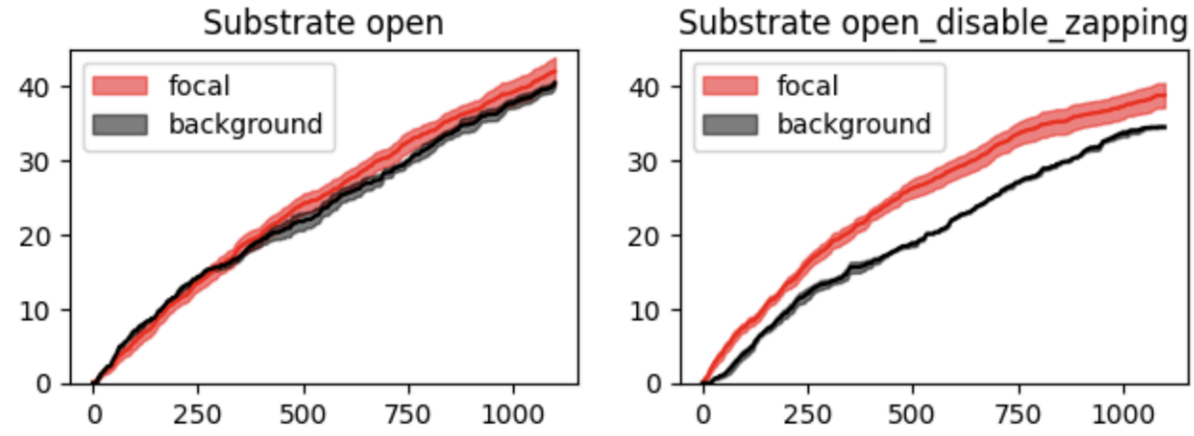


Fig3. No_zap agent focal population performed better when receiving a background population of agents that were able to zap and other agents that were not able to zap.

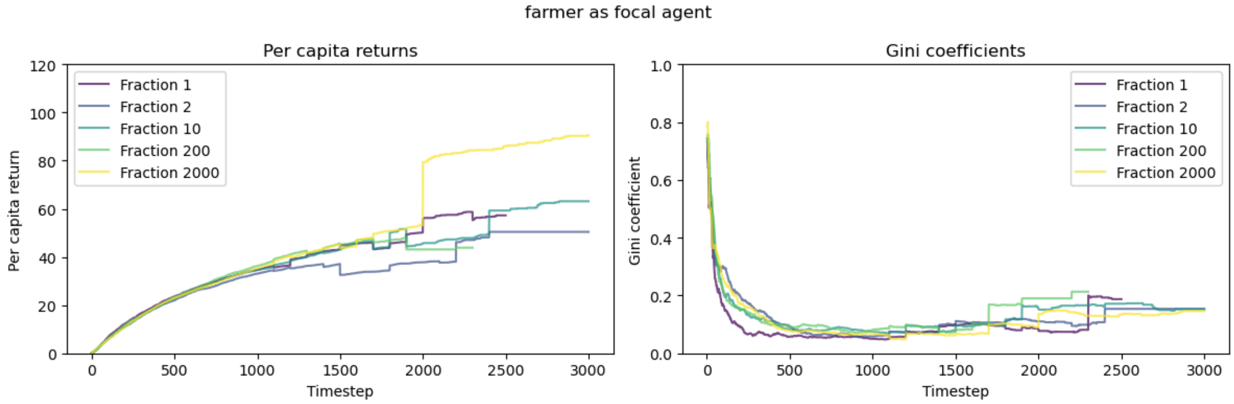
Varying Punishment Dynamics

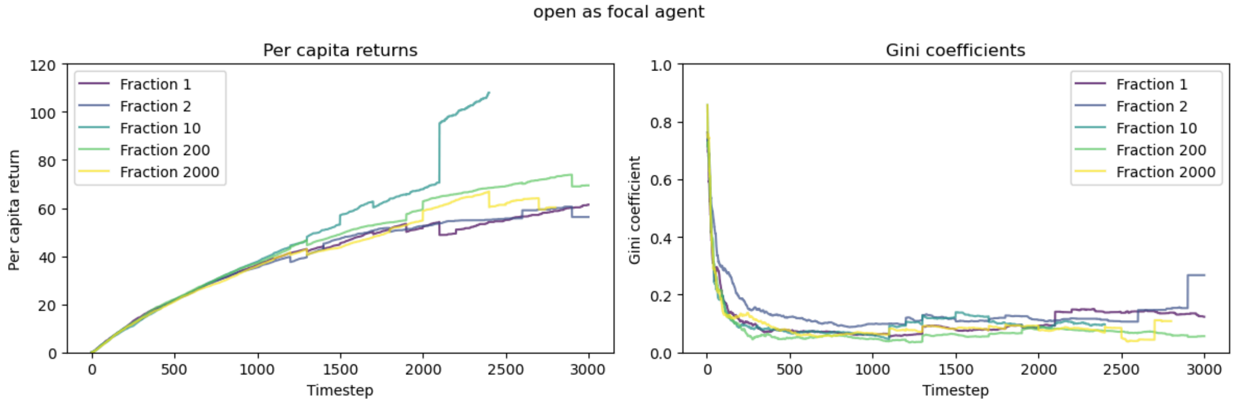
Beyond disabling zapping, we also investigate the effects of different zapping cooldowns on the agents' behavior in order to understand more about punishment dynamics. The Melting Pot framework allows for the change in Zapping Cooldowns, which corresponds to a step interval in which the agent has to wait before being able to zap again.

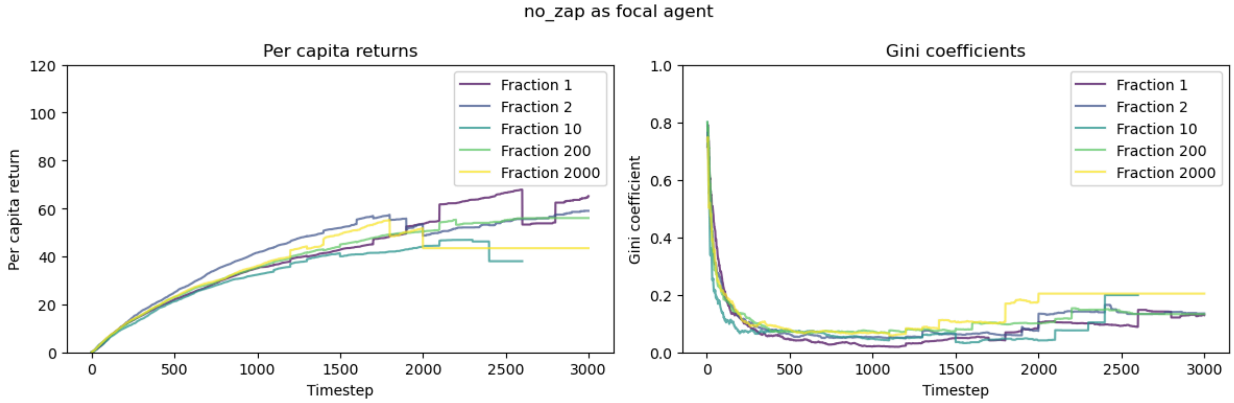
Each experiment was run over 15 episodes, on the Open Commons Harvest substrate with the baseline apple respawn rates. The selected Zapping Cooldown rates were 1, 2, 10, 200, 2000, which we refer to as "Fractions". A Cooldown rate of 10, for example, means that if an agent zaps another at a given moment, it can do so again only after 10 timesteps. A Cooldown rate of 1 is used as a control for agents that can zap at each timestep. In contrast, a rate of 2000 refers to a scenario of no zaps.

In this experimental setting, only the focal population was analyzed. The original background agents were trained by DeepMind over several more timesteps, and when placed together with our trained agents, would quickly deplete the apple orchard and impose an obstacle to our analysis over the focal population. Therefore, the substrate setting for this experiment does not have an initial threat to the sustainability of the rewards and isolates the behavior of the focal population according to different Fractions.

We examined the effects of the Fractions on the metrics of Per Capita Returns and Gini Coefficient for a selection of agents trained on the Farmer, Open, No Zap, Scarcity, and Private Property Substrates.


farmer as focal agent
Per capita returns
Per capita return
0
20
40
60
80
100
120
Timestep
0
500
1000
1500
2000
2500
3000
Fraction 1
Fraction 2
Fraction 10
Fraction 200
Fraction 2000
Gini coefficients
Gini coefficient
0.0
0.2
0.4
0.6
0.8
1.0
Timestep
0
500
1000
1500
2000
2500
3000
Fraction 1
Fraction 2
Fraction 10
Fraction 200
Fraction 2000


open as focal agent
Per capita returns
Per capita return
0
20
40
60
80
100
120
Timestep
0
500
1000
1500
2000
2500
3000
Fraction 1
Fraction 2
Fraction 10
Fraction 200
Fraction 2000
Gini coefficients
Gini coefficient
0.0
0.2
0.4
0.6
0.8
1.0
Timestep
0
500
1000
1500
2000
2500
3000
Fraction 1
Fraction 2
Fraction 10
Fraction 200
Fraction 2000


no_zap as focal agent
Per capita returns
Per capita return
0
20
40
60
80
100
120
Timestep
0
500
1000
1500
2000
2500
3000
Fraction 1
Fraction 2
Fraction 10
Fraction 200
Fraction 2000
Gini coefficients
Gini coefficient
0.0
0.2
0.4
0.6
0.8
1.0
Timestep
0
500
1000
1500
2000
2500
3000
Fraction 1
Fraction 2
Fraction 10
Fraction 200
Fraction 2000

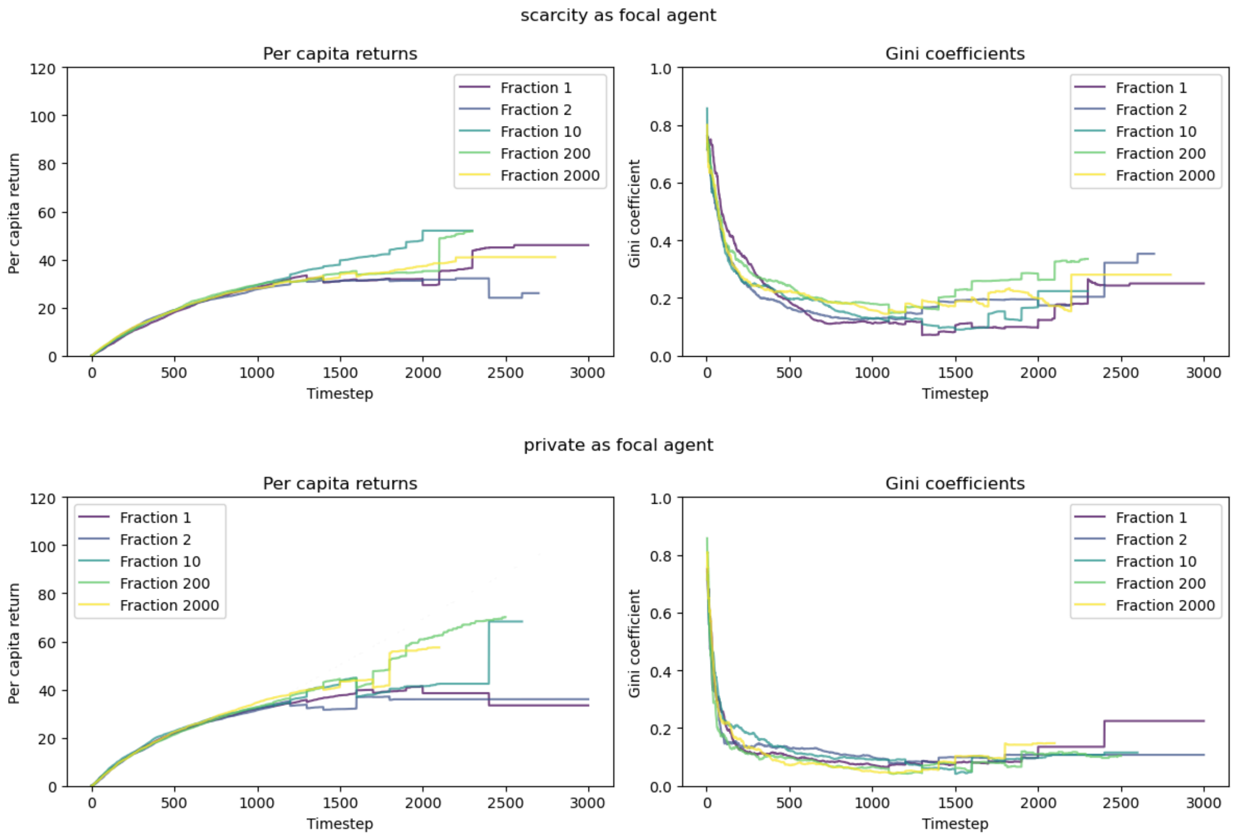


Fig4. Focal populations with various zapping cooldown dynamics

Important insights from the analysis include that Fractions of Zapping Cooldowns began to significantly affect overall returns after 1000 steps across all agents, with a particular highlight to Fraction 10 for the local agent.

In general, we noticed that the Gini coefficient across all agent settings remained consistently lower than 0.2 up until 1500 steps. However, after this mark, coefficients begin to diverge, with a particular highlight to the scarcity agent. The high variability in the Gini could indicate that as the episode progresses, the inequality of reward distributions among agents tends to increase, especially for Fraction 2. This suggests that scarcity agents become more competitive over time.

We also analyzed the percentage of Fully Depleted Simulations – substrates where the apples have been completely consumed – across different Fractions. Open agents demonstrated the lowest percentages of depleted simulations. This could indicate that the proposed alterations in the simulation environment made the focal population less effective in their resource management. The exception to this trend was the agent No Zap simulated on Fraction 2, which exhibits less depleted simulations than an Open agent simulated on the same fraction.

Farmer agents displayed the highest percentage of depleted simulations on Fraction 1, which contrasts with the low Gini curve presented in Figure 4. This could suggest that despite the rewards for ensuring the sustainability of the orchard, the ability to constantly zap each other could have led the agents to develop more **aggressive behavior** and thus become more competitive for the rewards. However, the fact that the Gini coefficient curve remains low compared to other fractions may suggest that all agents preferred to become more competitive at once and colluded to consume all the apples to achieve the rewards due to intense pressure from zapping. The agents opted to guarantee immediate rewards rather than to plan to collaborate for the sustainability of the orchard.

Another evidence of collusion can be noticed in the Gini coefficient spikes for the scarcity agents, especially after 1500 steps. However, instead of the collective decision to gather immediate rewards, as it happened to the Farmer agents, two or more Scarcity agents could have collaborated to collect apples at the detriment of others. In this case, the competitive pressure stems mainly from the scarcity of resources, rather than the zapping rate.

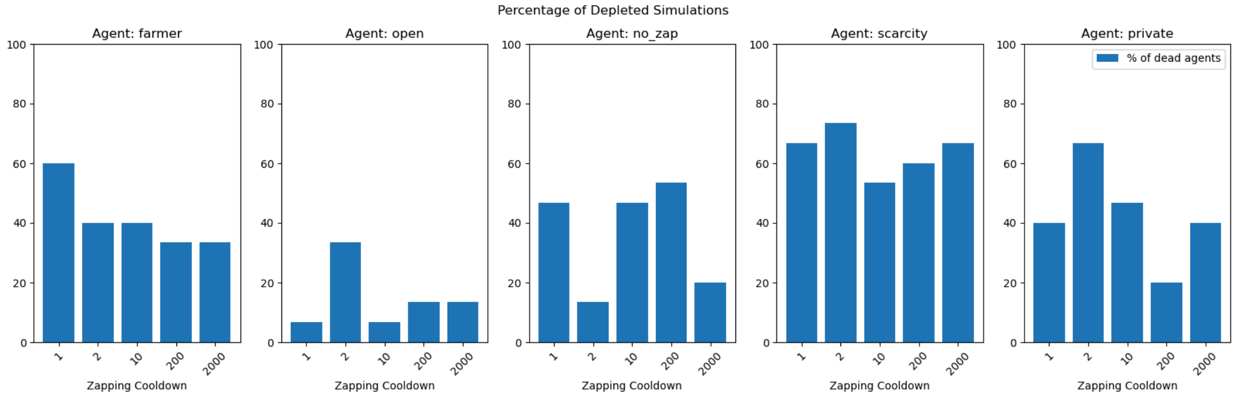


Fig5. Percentage of fully depleted Simulations across different fractions

Experiment 2 . Reward Mechanism

Implementing different reward mechanisms investigates how incentives influence agent behavior. From an AI safety standpoint, this experiment examines **the potential for misaligned incentives** leading to selfish behavior. Testing various reward structures, it ensures that agents are guided towards cooperative and fair resource management, minimizing risks associated with poorly designed incentives, as highlighted in the work on avoiding reward hacking.

The reward change dynamic that we propose is to reward the observer for observing the apple field next to them, so they can become vigilant about the resource depletion dynamics.

What we found in this case is that there was no substantial difference between the farmer agent and the rest of the agents, so changing reward design by an order of magnitude comes as a rational next step for the experiments.

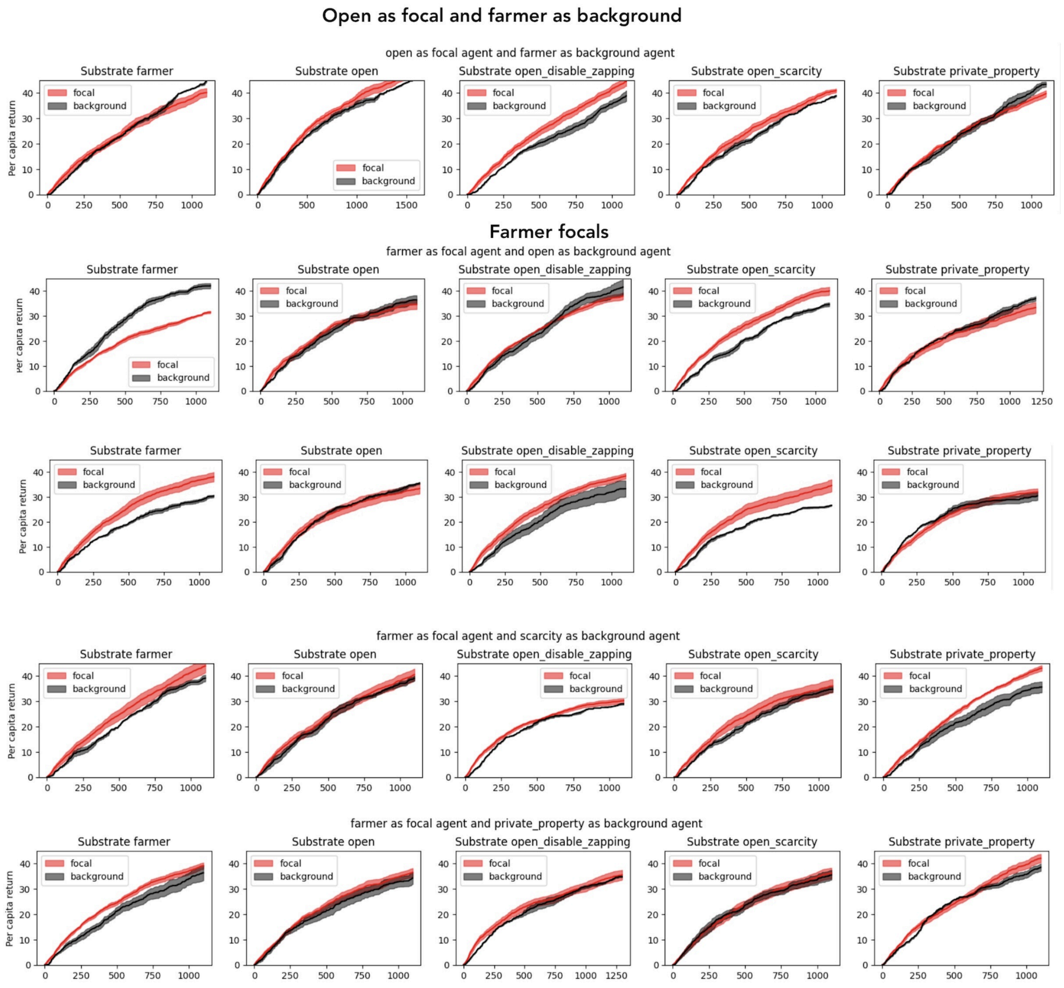


Figure 5. Comparison of focal vs background performance in farmer agent (trained with reward changes) and open baseline. The farmer population overall outperforms the background population, but there is no consistent symmetry when they act as a background population with open agents as focal, which has led us to propose increasing the number of episodes for evaluations

Experiment 3 . Environment Modification.

Modifying the environment tests the adaptability and robustness of agents ́strategies. This is crucial for AI Safety as it helps identify how changes in environmental structures impact resource management and whether agents can avoid over-exploitation in dynamic scenarios. This experiment helps in understanding the **resilience of AI** systems to environmental changes, ensuring sustainable resource use. This aligns with the concerns about distributional shifts and the robustness of agent strategies in diverse environments as discussed by Amodei et al.[22]

Experiment 4 . Resource respawn

Varying resource respawn rates evaluate an agent's ability to manage resources sustainably under different conditions. This experiment is important for AI Safety because it examines the **risks of overharvesting and the system's ability to maintain equilibrium**. It ensures that agents can adapt to changing resource availability without causing long-term depletion, a key aspect of sustainability.'

For this experiment, each agent was evaluated in an environment where the respawn rate was multiplied on a log scale ranging from 10^{-2} to 10^{2}, with specific intervals of 10^{-2}, $10^{-1.5}$, 10^{-1}, $10^{-0.5}$, 10^{0}, $10^{0.5}$, 10^{1}, $10^{1.5}$, and 10^{2}. A total of 15 evaluations were run for each experimental setup, with the averages being discussed below. Theoretically, this provides a scale ranging from extreme scarcity to extreme abundance with increased resolution on more moderate cases. The open substrate was used as it provides a straightforward, unaltered environment, ideal for examining how agents maintain a resource pool.

Fig 1: Per Capita Return Default Agent

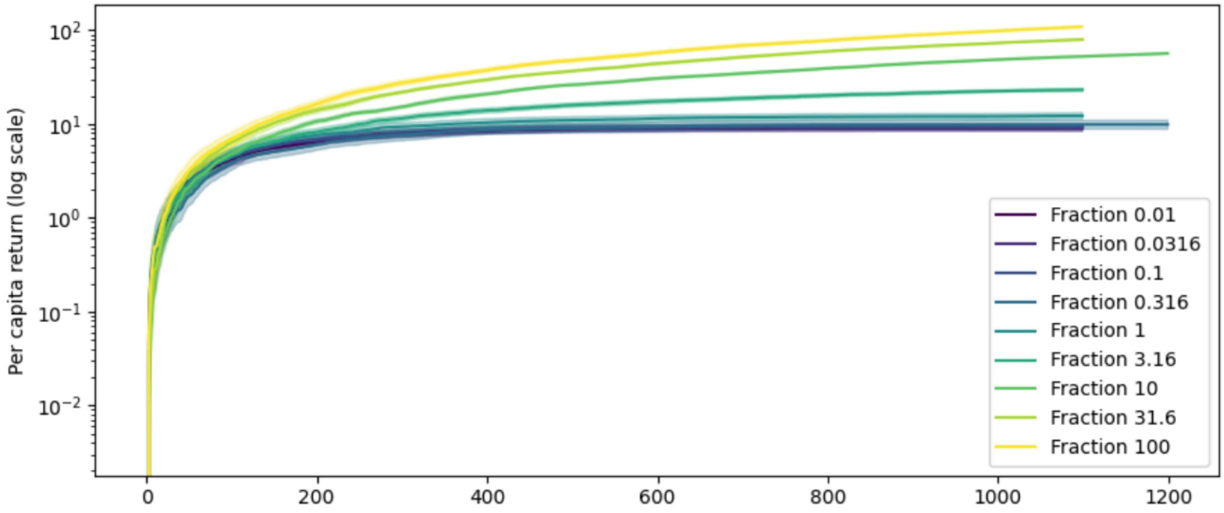


Figure 6 . The average over 15 episodes of the average per capita returns for the agents. This experiment was evaluated on the commons_harvest_open substrate with agents trained on the same substrate under varying degrees of abundance, as represented by fractions depicted in the key.

Figure 6 depicts the per capita return for the agent trained in the open substrate on the default regrowth settings. As expected, while the regrowth rate increases, agents generally see an increase in their per capita reward. Importantly, the graph shows that when agents are placed in environments with any decrease in abundance, fields are nearly completely depleted after only 200 time steps. This may imply that the agents have not learned the importance of depleting the last apple, but rather learned behaviors that approximate this goal, and break down once changes in respawn rates occur.

Fig 2: Equality Default Agent

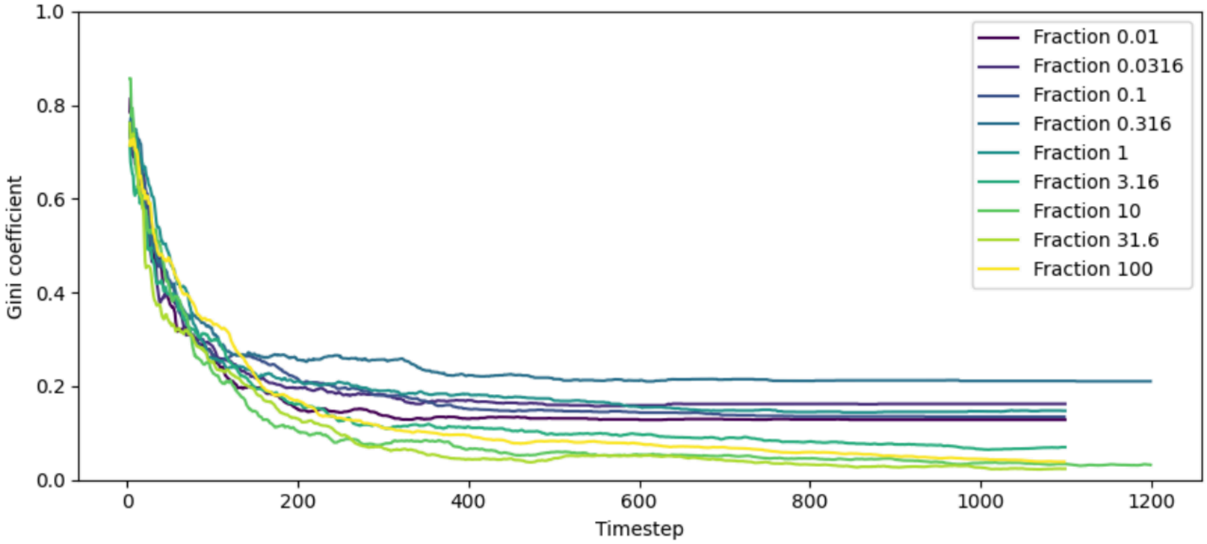


Fig 7. The average over 15 episodes Gini Coefficient for the agents. This experiment was evaluated on the commons_harvest_open substrate with agents trained on the same substrate under varying degrees of abundance, as represented by fractions depicted in the key.

The Gini Coefficient is a measure of inequality, where a coefficient of 1 represents a scenario where a single agent has the maximum reward where the others have none. A coefficient of 0 represents perfect equality among the agents. The graph depicts an initial rapid decrease in the coefficient across all levels of abundance, followed by a period of relative stability. At higher time intervals, a clear division occurs between agents in relative abundance (greens and yellows) and the agents in scarce environments (the blues and purples).

Fig 8. Fully Depleted Simulations

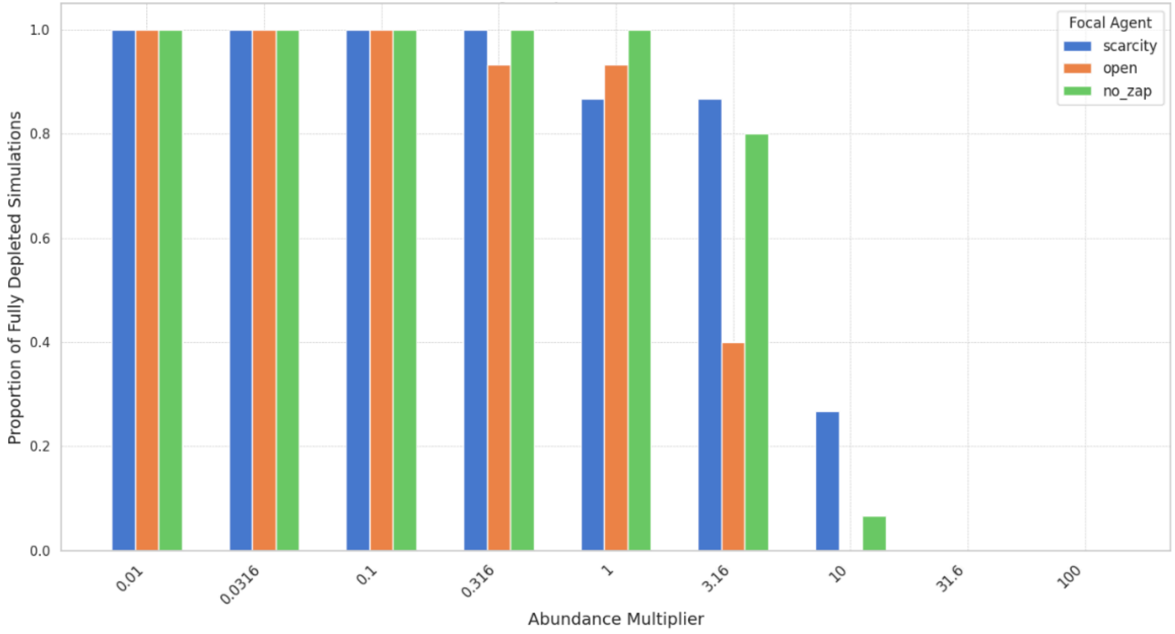


Fig 8. The proportion of the 15 simulations which have become fully depleted. Full depletion was determined by the lack of any increase in reward for 50 timesteps. The scarcity focal agent was trained on the commons_harvest_open environment with an abundance multiplier of 0.5. The no_zap agent was trained on the commons_havest_open environment with typical abundance but with zapping disabled.

In Fig 8 as expected, as abundance increases, the occurrence of fully depleted scenarios decreases for all substrates. However, we were surprised at just how poorly all of the agents generalized to less abundance environments. We hypothesized that the scarcity agent trained in a scarce environment would be more robust against decreasing levels of abundance. However, this was not the case. The scarcity agent performed *worse* than the open agent trained on normal levels of abundance.

These results could be due to our agents being poorly trained in general. As observed, the majority of evaluations resulted in fully depleted simulations even when being evaluated in their training environment. Ideally, we would be using agents that *very rarely* deplete the resources in the training environment, and then evaluate them on more scarce environments. This seems like a promising area for future work.

Safety discussion: Centralized Learning and Collusion in MARL

The basic approach in MARL is that of decentralized critique. During training, each agent updates their policies using their own critic computed from only their individual information. This reflects the realistic assumption that agents are fully independent and do not have access to the same information. One can expect that the lack of shared information is an obstacle for coordination. This motivates various centralized critique approaches that modify the training procedure to include shared information in some way. The intuitive rationale for this is that using shared information can allow training to capture the interdependencies between agents' actions, leading to more coordinated and efficient strategies.

Centralized critique comes in different flavors. On the one side, we have a centralized policy approach, where a single policy is being trained for all agents. This solution treats the multi-agent setting effectively as a single-agent one. Hence, it may be seen as inadequate as our practical interest is in agents that are independent, at least to some degree. That being said, centralized policy is definitely of theoretical interest as it gives an upper bound on performance of decentralized policies and hence, it can be used as an effective baseline.

Between decentralized critique and centralized policy lie different methods which aim at training multiple decentralized policies using some form of centralized critic. During training, each agent updates its own policy but does it using shared information about observations made by other agents. The shared information appears only during training and not in evaluation. More formally, this means that appropriate policy gradients are estimated using joint value functions. A paper by Lyu et al [23] references three specific algorithms, where expected joint performance is conditioned on shared information about history, about state or about both history and state.

One could argue that this decentralized policy with centralized critic is not much more realistic than the simple centralized policy. Indeed, if our goal is to study how coordination can arise between independent agents, should we bother with a scenario where somehow each agent knows the memory of other agents? A possible answer to that question could be similar as in the case of centralized policy: maybe the mixed approach can still serve as a benchmark for performance?

Lyu et al [23] provided some insights into why such a view may be too simplified, if not simply false. In particular, one of their theoretical findings is that state-based centralized critics may increase bias, while the policy gradient variance of centralized critics is at least as large as that of decentralized critics.

Here is something about how centralized criticism relates to multi-polar alignment failure.

Research Literature [24] claims that there exist misconceptions regarding centralized critics in the current literature and shows that the centralized critic design is not strictly beneficial, but rather both centralized and decentralized critics have different pros and cons that should be taken into account by algorithmic designers.

Conclusion

An important consideration with multi-agent systems is the risk of collusion. **Collusion is generally defined as 2 or more agents (covertly) coordinating to the disadvantage of other agents**. Many multi-agent games (including most of the ones described in the Melting Pot framework) prescribe fully independent learning and preclude communication between the agents, so in such scenarios, collusion is explicitly disabled. However, in open-world settings without artificial constraints, attempting to study and mitigate this phenomenon is critical to ensure safety.

In their paper[25], Foxabbott et al define and propose interventions to mitigate collusion within the context of partially-observable stochastic games (POSGs), which is said to be a general model for real-world multi-agent AI systems. While collusion is commonly assumed to be covert and intentional, their definition is agnostic to both of these aspects and instead only focuses on mutual benefit for the colluding agents at the expense of others. They specify three types of interventions to prevent rational agents from employing such strategies (some of which we have also tested in our experiments):

1) Adding **noise to observations** (similar to real-world imperfect information) - We did not attempt to modify agent observation spaces as part of this set of experiments.
2) **Restricting agents' action spaces** (similar to real-world regulation) - We attempted variations of this in our 'no_zap' and 'varying punishment' experiments by limiting or disabling the agents ability to zap others. In the 'varying punishment' experiments where we modified zap cooldown times, we noticed that the Gini coefficients initially dropped quickly with timesteps and reached a minimum of around 0.1-0.15 by 500 timesteps (see figure 4 above). However as apple fields start getting depleted, we often saw an increase in the Gini coefficient from approximately 1500 timesteps onwards, indicating potential collusion among sets of agents. This was especially pronounced in the scarcity-trained focal agents, suggesting that environmental pressures could exacerbate the likelihood of collusion.
3) Modifying **agents' payoffs** (similar to changing incentive structures to shape behavior) - One method of updating the reward mechanism we experimented with was by training Farmer agents. In this case, the agents got a reward for when apples were present in their observation space, thus hopefully encouraging them to conserve and promote regrowth of apples. However our experiments comparing farmer agents to the baseline were inconclusive, possibly due to a low magnitude of observation-based reward.

Another domain-independent alternative to detecting collusion in multi-agent systems using an information-theoretic approach has been described in a paper by Bonjour et al[26]. Here they propose a ***collusion-detecting algorithm*** by generating and analyzing a joint policy matrix based on the outcomes of either partially observable sequential games or repeated fully observable simultaneous games, and then comparing the pair-wise net influences to a collusion threshold.