The Effect of Centralized Group Trust on Reinforcement Learners Observed Using the "Stag Hunt" Model

Benjamin Ellison

Student ID: 720052156

Abstract

Trust is a social asset which is foundational to human life, and exists at several scales. While trust between singular individuals is not too difficult to quantify, when the scale is broadened to a societal level, the mechanisms of trust become much less trivial to evaluate and understand. Reinforcement learning (RL) models provide a very promising opportunity for insight in this respect, having been used in the past to model bilateral trust using the framework of standard game theory contexts to great success. Despite this, the role of broader, sociological reputation systems is an area where study remains largely unexplored, possibly owing to the complexities of defining this behaviour, or recreating it in a controlled environment. As well as attempting to rigorously define and model these behaviours, this report seeks to evaluate how these systems may benefit or hinder small groups of Agents, using RL models which navigate Stag Hunt scenarios with minor variations. These Agents consider how bilateral trust as well as group-wide reputations, and how these interacting behaviours impact efficiency, both on an individual and global level, will be analysed. This is with the dual hopes of gaining a better understanding of the function of each form of trust and finding out how each may be used to aid in the efficiency of future cooperative RL models.

Declaration

I acknowledge the following uses of GenAI tools in this assessment:

□ I have used GenAI tools to:

- □ develop ideas.
- □ assist with research or gathering information.
- □ help me understand key theories and concepts.
- □ identify trends and themes as part of my data analysis.
- □ suggest a plan or structure for my assessment.
- □ give me feedback on a draft.
- □ generate images, figures or diagrams.
- □ proofread and correct grammar or spelling errors.
- □ generate citations or references.
- □ Other: [please specify]

□ [X] I have not used any GenAI tools in preparing this assessment.

I declare that I have referenced all use of GenAI outputs within my assessment in line with the University referencing guidelines.

I certify that all material in this dissertation which is not my own has been identified.

Signature: Benjamin Ellison

Contents

1 Introduction

Human interaction is a vital part of all areas of study. Arguably, it is the lens through which all knowledge is gathered, since all information accumulated is judged, in part, by its source. In spite of this, our understanding of the manner by which trust is formed and utilised is still, largely, minimal. This report seeks to illuminate a dichotomy between types of trust, and examine why this distinction may have merit, and what the function of each type is.

The two types of trust examined in this report are direct reciprocal trust and indirect reciprocal trust. Direct reciprocal trust being that which an Agent builds based on prior experience with some specific other entity, whereas its counterpart is built based upon the reporting of a community of other trusted Agents, taking into account the trustworthiness of the reporter, as agreed upon by said community (Korsgaard 2018). For the sake of convenience, throughout the remainder of this document, indirect reciprocal trust will be called 'Reputation', and its counterpart simply 'Trust'. These two categories reflect the dichotomy of primary and secondary information sources, as well as the burden of reputation placed upon information by its conveyor and their community (Barclay 2015). It could also be seen to reflect the more sociological phenomenon of the forming of closely-knit societal out-groups (Pan and Houser 2013). The term used to encompass trust as a concept, being a quantitative proclivity for one given Agent to cooperate with another, will be referred to as 'Faith'.

Indeed, it would seem Reputation comes with a much higher degree of complexity than the relatively trivial Trust, since several sources must be accounted for, and allows for the potential for anomalous mistakes to proliferate across communities, tarnishing the Reputation of an Agent which may otherwise be a valuable partner for cooperation. This additional complexity alongside high risk of misjudgement would, at face value, seem to suggest that Reputational systems are ineffective as means of maximising educated choice in collaborative partner, and yet it seems present, if not encouraged in so many areas of life. It exists in social media, with the prominence of quantified interaction metrics (Liu and Sun 2014), and guides our financial interactions with reviews and recommendations (Resnick et al. 2000). This would imply that there is some gain to be had in this form of many-to-many, group feedback system.

RL systems seem uniquely equipped to provide insight into this line of inquiry, since they provide us something no level of ethical human study could provide: a fresh, relatively unbiased starting point. This has an opportunity to be invaluable, because it removes the well-known biases (e.g. race, gender, age) that human cooperative partners have, allowing for the problem to be examined in a more refined way (Ruhl 2023). Another benefit is the ease of empirically quantifying these interactions and how they develop over time without relying on human surveys, which are known to be a potentially dubious means of collecting robust data (Kohler 2023). The associated drawbacks of these, however, is that it's likely the results of this study may not lead to actionable conclusions that will lead to predictable beneficial outcomes when applied to human cooperative systems, since cognitive bias is an impossible trait to act purely outside of, beyond enforced ignorance, which discounts interpersonal trust altogether.

Nevertheless, with an increasing push for automation across industries, this study may still offer a valuable insight into how best to structure cooperative RL systems, whether or not they ought to have a shared knowledge base, and what potential gains and losses the proposed system provides. It is the intention of this study to ascertain the nature of these benefits, and how they may be usefully implemented in future cooperative group RL systems.

2 Specification

A system by which Trust and Reputation are to be examined requires a number of prerequisites:

- A test of Faith (an exercise between two Agents which jeopardizes their trust and reputation values, but also offers a chance improve them)

- Agents which use RL both to engage with the test and to select partners
- A means by which Trust data can be stored and accessed
- A means by which Reputation data can be stored and accessed

2.1 Stag Hunt (A Test of Faith)

There are several means by which Agents can engage with a test of faith. These however, can vary in output types and nature quite drastically, some lending themselves better to a study on Trust and Reputation better than others. In order to draw from well recognised and standard procedures, the doctrine of game theory was drawn upon. Game theory provides a litany of abstracted archetypes for tests of faith (henceforth referred to as 'games'), each of which has its own benefits and shortcomings. This pool of candidates can be refined by eliminating games which necessitate an extensive component of randomness, and games which are unambiguously non-cooperative. The two most simple archetypes remaining were the Prisoner's Dilemma (Chen 2024) and the Stag Hunt (Ahlstrom 2023). These games are identical in terms of choices for the state space, but very distinct in terms of outcome. Where, in the prisoner's dilemma, a given Agent benefits by acting in a manner which seeks to harm their partner, Stag Hunt offers no such incentive. The Prisoner's Dilemma is an exercise in strategic betrayal, wherein an Agent is most rewarded when trust is built by sacrificing long term gain in the interest of conditioning their partner to act in a way that rewards both Agents a small amount, only to shake this strategy later, hoping to do so before the partner does. In contrast to this, Stag Hunt is an exercise in strategic inter-Agent risk management, where the optimal outcome is symmetrically optimal for both Agents. This makes Stag Hunt a better candidate for pure Trust and Reputation analysis, since all Agents will tend toward the same outcome, building Faith, and a breakdown of Faith is disincentivised.

The primary detriment the Stag Hunt model suffers, in the context of compatibility with RL models, is that it is quite trivial. A given Agent's convergence on the optimal outcome depends primarily on their knowledge of the other, and the options afforded to it, and once an Agent can be relatively confident that the other will take the large reward, it is unlikely to change from this Nash equilibrium for any reason beyond the typical rate of exploration present in such models. This is excellent for use as a control test, since Agents functional effectiveness can easily be examined in contexts where the ideal outcome is static, and whether this outcome is attained, the time spent doing so, and behaviour therein is all that can be studied, but in order to test the robustness of a Reputation system, something more complex seems sensible, and this can be introduced in the form of secondary games. Such as to minimize distinction between tests, two other games are introduced to the study, each variations on the Stag Hunt game, but which tweaks to the rewards such that their equilibria are meaningfully distinct.

2.2 Variations on Stag Hunt

2.2.1 Bison Hunt

Bison Hunt is a game created for this study. It is distinct from Stag Hunt in that the Stag is replaced with the Bison. The Bison starts with a higher reward as the Stag, but as the study continues, it decreases according to some logarithmic function (described in section 3.2). This leads to a gradual decrease in value, such that, over a certain number of Hunts, the value will decrease to a point wherein the value of the Bison is less than the value of a Hare, meaning that one of the two Nash equilibria the typical Stag Hunt has is a false flag, necessitating a swap in tactics in order to maintain optimality. This will mean that Agents must relearn their tactics in order to optimise score, navigating a linearly Faith-straining context.

2.2.2 Random Hunt

The manner by which Random Hunt differs from Stag Hunt is that the reward function replacing the Stag's static high reward is a randomly chosen number. This means that Agents will consistently have their trust formed partially on unrelated intel based on random outcomes. These random scores will range anywhere from just slightly higher than agreeing on the Hare to higher than the Stag in the original Stag Hunt. Crucially however, between these two values lies the reward that the Agent that chooses to Hunt the Hare receives when their partner doesn't, meaning that the choice is more between stability and probabilistically determined risk.

2.3 RL-Powered Agents

The Agents who participate in these games will have two choices to make per cycle. The first is partner selection, which will take into account their own personal Trust in each candidate and the Reputation of said candidate within their community. Once they are paired with an Agent, they must decide the best course of action to take, given the Agent they are with. This will be based entirely on their past experiences with this partner. In order to give Agents a meaningful distinction between choice of partner, it is advisable that some Agents should have biases toward certain behaviours, perhaps some prioritising exploration while others lean more toward exploitation.

2.4 Trust and Reputation Systems

The systems by which Trust and Reputation are stored should vary primarily in where they are stored. Where Trust tables are private, and unique to each Agent, the Reputation table should be public and visible to every Agent. After each Hunt, every Agent should update their partner's entry on their Trust table, and on the Reputation table.

3 Design and Development

This experiment took the format of a Python program which ran simulations of the Hunts described in Section 2.2 and collated the data into the Figures analysed in Section 4. This section will describe the parameters of these simulations, as will as the specific nature of the functions used, and the justification for these choices.

3.1 System Outline

The tests run to ascertain metrics are structured as follows:

1. 8 Agents select their preferred partners based on trust and reputation scores
2. The Agent pairs play the given "Stag Hunt" variant game 20 times
3. Trust and Reputation scores are updated based on the results of the test
4. Repeat steps 1-3 1000 times
5. Repeat steps 1-4 10 times for each Hunt type with a new set of Agents, then take an average of the total group reward
6. Repeat steps 1-5 once per Reputation multiplier (henceforth called "Gossip Value")
7. Save results

These values were tuned in the interest of taking the minimum relevant data surrounding convergence. Tests were also run wherein step 4 contained more repetitions, but these yielded tests wherein convergence had been achieved, and more training yielded negligible results.

The remainder of this section will go into further detail regarding the minutiae of each step, describing and explaining the involved mechanics.

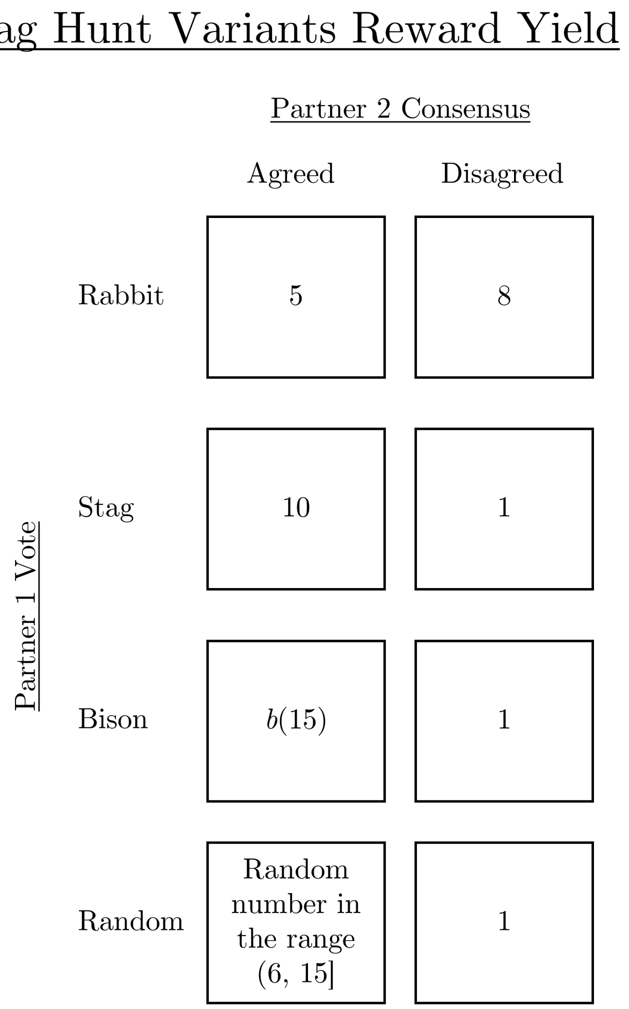
3.2 Hunts

The Hunts used in this experiment (Stag, Bison and Random) yield the following results (see Fig.1.)

Bison Hunt uses the function b to steadily decay the reward yield. The definition of the function b is as follows: $b(r) = \frac{r}{1.0001^x}$

Where:

- x is the number of times the function b has been called.



Stag Hunt Variants Reward Yield

		Partner 2 Consensus	
		Agreed	Disagreed
Partner 1 Vote	Rabbit	5	8
	Stag	10	1
	Bison	$b(15)$	1
	Random	Random number in the range (6, 15]	1

Figure 1: The reward yields offered by each Stag Hunt variant (each score represents the reward Partner 1 would receive in the described interaction. Partners are interchangeable)

3.3 RL Agents

The Agents in this experiment use Epsilon-greedy Q-learning to govern their decision-making processes. This is a suitable method for a number of reasons. Epsilon-greedy learning allows for direct examination of decision-making, since it does not use any Deep Neural Network components. Its entirely probabilistic method of triggering explorative behaviours, while a simplification of real-world decision-making processes, allows for a suitable simulation of these behaviours when applied to a task that is both trivial and reasonably small in scale. Since the decision spaces are either the number of Agents n, or the number of Hunt options 2, both tasks for which Q-learning is used are suitably trivial to avoid the use of any more convoluted RL method. Epsilon-greedy decision-making also means that there need only be 2 behaviour types: explorative and exploitative, which makes the outcomes of each behaviour easy to monitor and model.

As mentioned, the Agents used in these simulations have 2 phases of decision making: Partnering and Hunting. These are separate since candidate partner preference and actual partner behavioural prediction are unrelated, being where Trust is used and built, respectively. The functionality and mechanics of each behaviour will be elaborated on for the remainder of this section

3.3.1 Partnering

Partnering is when an Agent will request that some specific other Agent be its partner for the next 20 Hunts. If the Agent is being exploitative rather than explorative, it will consult its own Trust table for each candidate Agent (which will have collated associated reputational data, see Section 3.4) in order to choose, via cumulative probability, its preferred candidate partner. Otherwise, if the Agent is being explorative, it will chose a candidate partner completely at random.

The function used to select a partner, assuming exploitation, is as follows:

$$c = rand(0, 1)$$

$$C = cdf(g \times scale(R_{available}) + scale(T_{available}))$$

$$partner = pos(min([z : z < c | z \in C]))$$

Where:

- *scale* is the Min-Max Feature Scaling algorithm
- $R_{available}$ is an array containing the Reputation values of available partner candidates
- $T_{available}$ is an array containing the Trust values of available partner candidates
- g is the current Gossip score, scaling by a factor of 0, 1, 2, 5, 10 or 100
- *cdf* is the function mapping the resulting array to a cumulative distribution
- $rand(x, y)$ is a pseudo-random function which finds a float in the range (x,y)
- *pos* is a function which maps the position of the value in the array to the corresponding candidate.

Where Agents have both selected one another to be partners, they will be paired together and removed from the pool of candidates. Among the remaining candidates, partner selection will then be repeated until one of two conditions is met: the first being that all partners are paired, the second being that no candidates are in agreement whereupon, each will be assigned a partner randomly. This aims to incentivise accord between partner candidates.

3.3.2 Hunting

Hunting is a more simple step, wherein, if the Agent is exploitative, the Hunt choice which has yielded the most positive results when paired with this partner in the past will be chosen and if they are explorative, the opposite is true. The result of both partners selected Hunt choices will incur the relevant reward (see Fig. 3.3.3).

3.3.3 Reward

Reward is granted based on the Hunt outcomes (See Fig. 1). Once reward is granted, the Agents take this value and apply it to both their partner selection table and their current partner's optimal behaviour table. They apply it using the following calculation (Karunakaran 2020):

$$Q(P, A) \leftarrow \alpha r + \gamma max(r + Q(P, A)) - Q(P, A)$$

Where:

- P is the current Partner
- A is the chosen Action
- $Q(P, A)$ is the Q value for action A) with partner P
- α is the learning rate
- γ is the discount rate
- r is associated reward for the Hunt outcome

This is a simplified version of a typical Temporal Difference equation adapted to the fact that there are only 2 choices and the progression of states is linearly controlled by the environment in a manner which is not entirely controlled by the Agent, as opposed to, say, Grid World traversal (Antony, Roy, and Bi 2023), wherein state management is the primary task for an Agent to do.

3.3.4 Parameters and Variations

In order to give Agents some reason to trust some partners more than others, 3 variants of Agents have been implemented, varying in readiness to explore suboptimal paths. The 3 variants and their values are as follows:

- Average Agents
  - Exploration Rate: 0.1
- Risky Agents
  - Exploration Rate: 0.3
- Hare-Brained Agent
  - Exploration Rate: 0.1
  - Always chooses Hare
- Hareless Agent
  - Exploration Rate: 0.1
  - Never chooses Hare

There are twice as many average Agents as there are risky Agents, and there is only ever one of each of the other two types, meaning that, in a group of 8 Agents, there are 4 average, 2 risky, 1 Hare-Brained and 1 Hareless. Shared between the 3 Agent types, however are the following parameters:

- Learning Rate: 0.1
- Partner discount Factor: 0.3
- Voting discount Factor: 0.9

These parameters have been tuned via repeated experimentation so as to reach an acceptable convergence in an acceptable length of time.

3.3.5 Epsilon Decay

In order to aid in stability of convergence, Epsilon decay is implemented such that Agents' exploration rate decreases throughout the experiment (Bowyer 2022). Rather, however, than steadily decreasing throughout the experiment, in the interest of not outpacing dynamic environments, this decay is dependent on a specific condition: All Agents match without requiring random pairing (as described in section 3.3.1). When this happens, all Agents' exploration rates are reduced by means of dividing them all by 1.001.

3.4 Trust and Reputation Modelling

Trust and Reputation are modelled as arrays. Each Agent will have an array of trust scores corresponding to each other Agent. With each Hunt, the value corresponding to their current partner is updated using the formula defined in Section 3.3.3.

Reputation is modelled similarly, with each Agent in a pair updating their partner's Q-score on a shared Q-table according to the following calculation:

$$R_{a_t+1} \leftarrow R_{a_t} + \gamma b_t * (r + \alpha * max(R_t) - R_{a_t})$$

Where:

- R_a is the current Partner's reputation score
- α is the environmental discount factor, which in this experiment is 0.8
- γ_b is the current Reporter's Learning Multiplier (Reputations are Feature Scaled to be between 0 and 1, and the Reporter's corresponding value is the Learning Multiplier used)

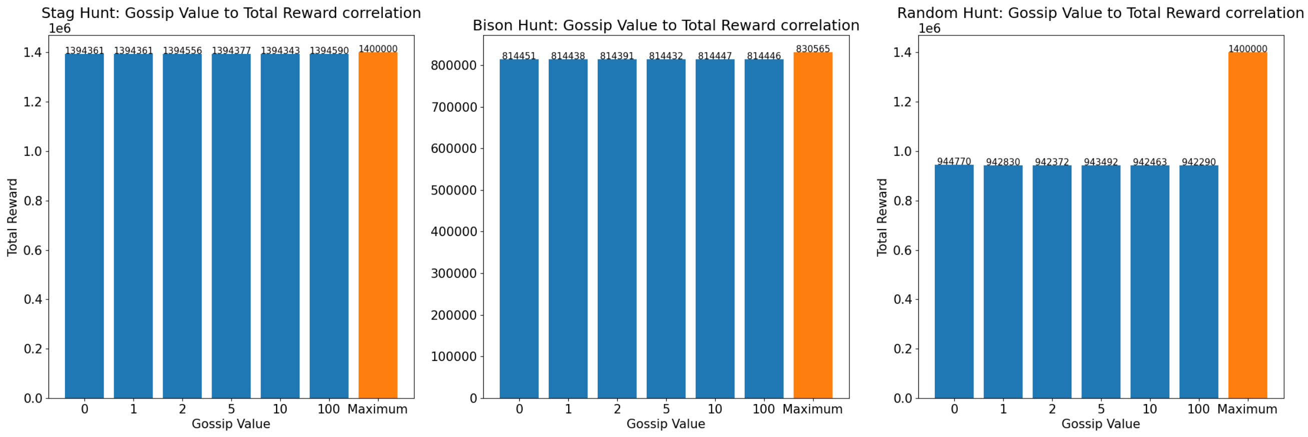
After these values are updated by every Agent, reputations are normalised, such that more recent actions impact an Agent's reputation more significantly, and the values do not scale excessively.

4 Results

The ensuing section will contain the results of the experiment. Each metric taken will have its own subsection, and within these subsections, where relevant, models will be viewed in terms of changes between Hunt types, Gossip levels, and Agent types. These results will be analysed in the following section, and conclusions drawn from the data will be discussed. Worth noting is that, For the remainder of this report, Agents will be referred to by number. The Agents have the following behaviours:

- Agents 0 and 1 are Risky Agents
- Agents 2 to 5 are Average Agents
- Agent 6 is the Hareless Agent
- Agent 7 is the Hare-Brained Agent

It is also worth noting that, for the vast majority of these results, to what extent a reputation system is used makes little difference to the outcome of Agent policy convergence in the scale examined. In the interest of thorough examination, results that are not intrinsically related to reputation will be shown in the data for the submission. **??**.



(a) Stag Hunt (b) Bison Hunt (c) Random Hunt

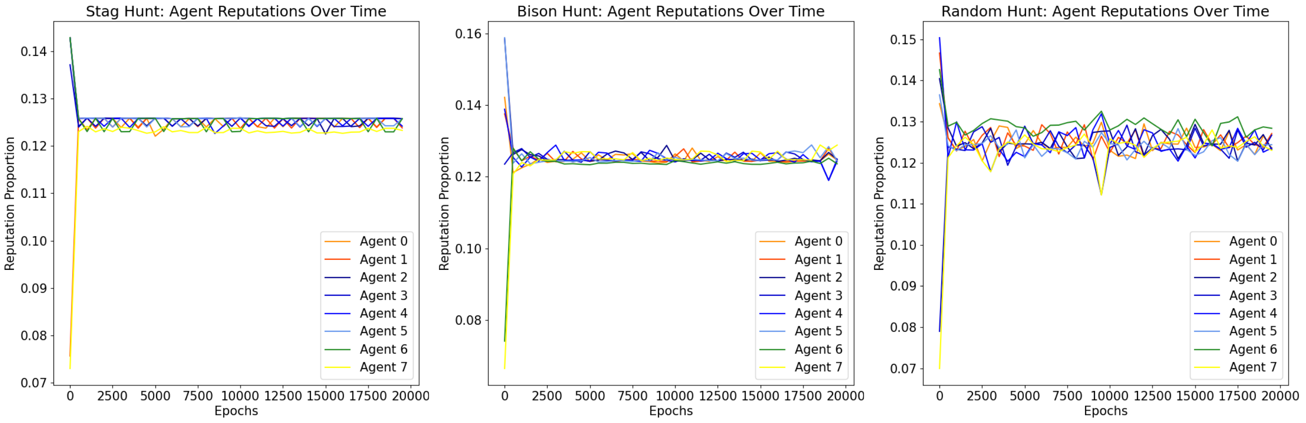
Figure 2: Effect of Reputation Multiplier on Total Group Reward

4.1 Total Group Reward by Gossip Level

The above data shows the total reward of each group for each simulation, averaged over 10 simulations, in comparison to the maximum theoretical reward possible, (given the Hare-Brained and Hareless Agents). In Fig **??**, it's worth bearing in mind that the pseudo-random number output is taken to be the median possible value (10), since it can be safely asserted that all values gained will trend towards it, thus why the maximum value is not the theoretical maximum, but the theoretical attainable maximum value, had the Agents chosen the optimal policy.

The two notable observations made from all three graphs is that the gains and/or losses to be made from the implementation of the given Reputation system are negligible, and that, with both systems, Agents are well adapted to both Stag Hunt and Bison Hunt, but struggle to adapt to Random Hunt.

4.2 Reputations Over Time



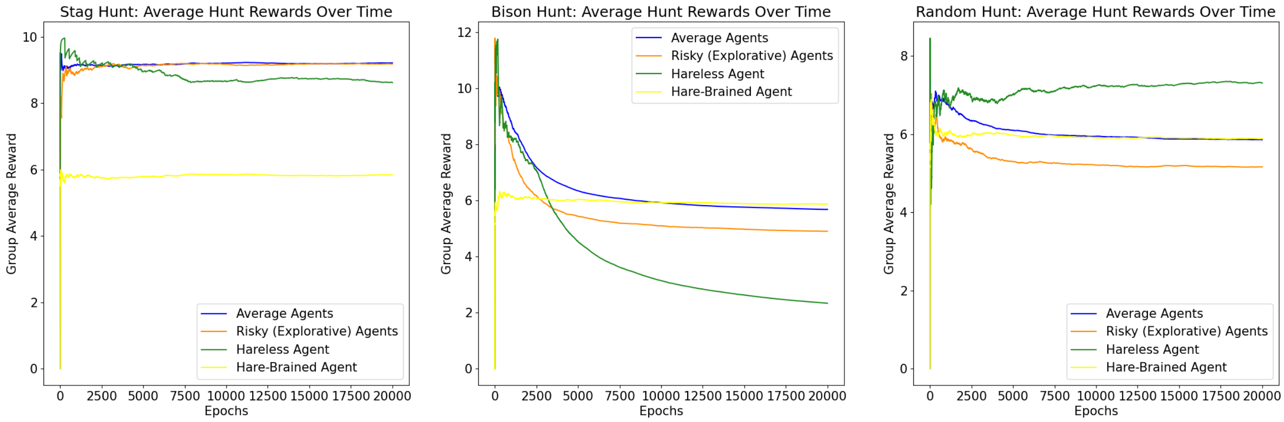
(a) Stag Hunt (b) Bison Hunt (c) Random Hunt

Figure 3: Development of Reputation Proportions for Each Agent Over Time

The graphs in Figure 3 show that, after the very beginning, where reputations are still being scaled to negate time bias, some Agents (primarily the highest scorers) maintain the highest reputation. This means that, in systems where reputation is the strongest consideration, the highest scorers are more likely to be paired with their ideal partner, since that candidate in turn will be trying to pair with them. This would, in theory, mean that situations wherein reputation is a stronger consideration would lead to higher polarisation, since all Agents would be vying to pair with the highest scorers, however, in action, the opposite becomes true, since when two high-reputation Agents co-operate, should any disappointment arise in terms of expected reward, the Agents' reputations drop rapidly and significantly, as can be seen in the Hareless Agent 6 in Figure 3a.

Predictably, Random Hunt varies more significantly in terms of reputation, but instead of the highest-scorer dropping in reputation severely, as is seen in the other graphs, the Hareless Agent 6 remains the Agent with the highest reputation. This may be because this Agent is easy both to cooperate with for an unpredictable positive outcome, or exploit for a more predictable, but still positive outcome.

4.3 Reward curve by Agent Type



(a) Stag Hunt

(b) Bison Hunt

(c) Random Hunt

Figure 4: Learning Curve Averaged for Each Agent Type

From these graphs, a number of assessments can be made regarding Agent behaviour. What they have in common is that they display a very fast convergence. In Figure 4a, this convergence is demonstrative of a suitable, settled policy, showing that the Agents very quickly gain an understanding of one another, as well as their environments, when the returns are static and predictable.

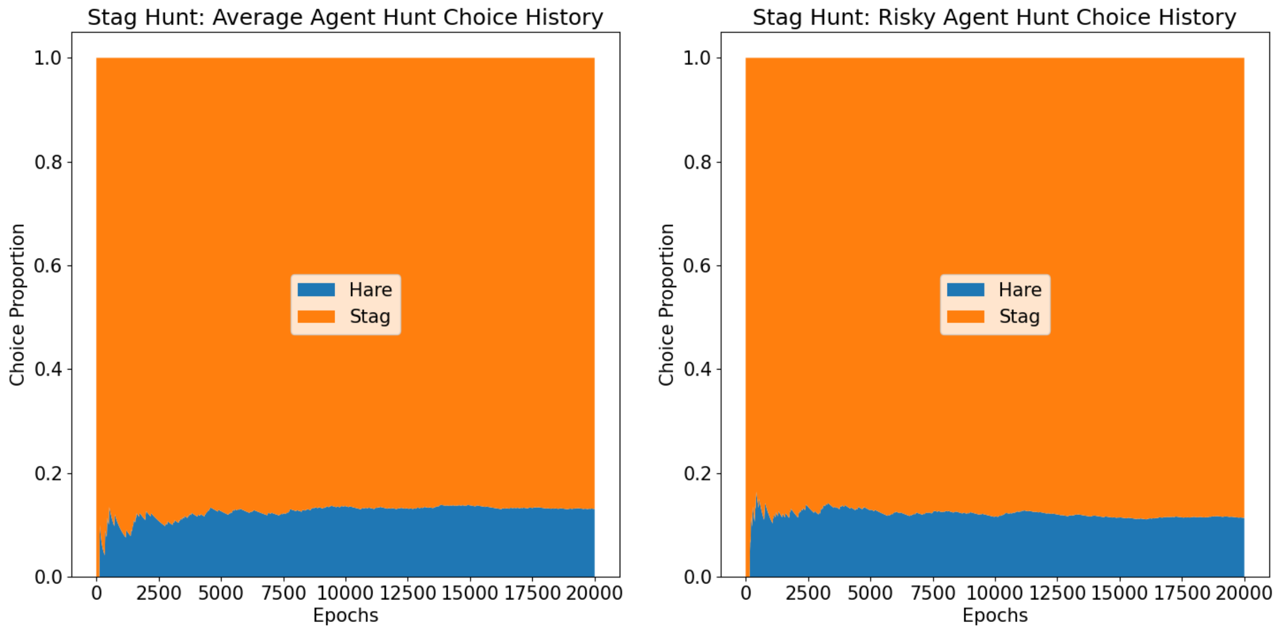
Figure 4b demonstrates that, even with Epsilon decay, Agents can still adapt very effectively to a slowly changing environment. Also observable in this graph is that Agents can update their policy very quickly when returns have diminished beyond reasonable continuation of their exploitative policy, as seen in the sudden drop in reward for the Hareless Agent at around 2500 epochs.

Figure 4c, on the other hand, show an inherent risk aversion to some Agents, as despite the random reward for a Random agreement averaging to more than even the most rewarding Hare outcome, Agents are still converging primarily towards voting Hare. This bias towards risk mitigation is not universal however, with Agents choosing between Random and Hare far more evenly throughout the simulation than was the case for Stag in the Stag Hunt simulations (as seen in the Fig 4a).

4.4 Votes per Agent Type

In this section, stackplots will be displayed showing the vote proportion of a randomly selected Agent of each type (note that only Agents whose choices vary will be shown; Hareless Agents never vote Hare and Hare-brained Agents always do) varied by Hunt type over time. Reputation is not factored into which graphs are displayed as this does not change the outcome in any way (shown in the submission's data).

4.4.1 Stag Hunt


Stag Hunt: Average Agent Hunt Choice History

Stag Hunt: Risky Agent Hunt Choice History

(a) Average Agent Hunt choices over time

(b) Risky Agent Hunt choices over time

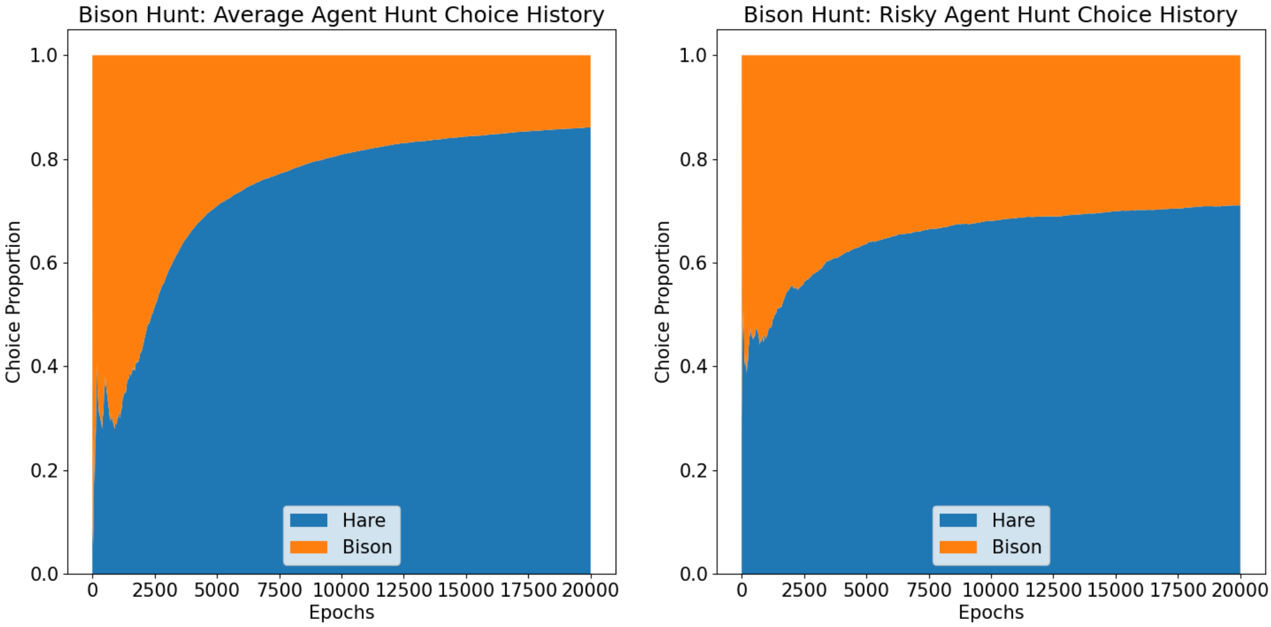
Figure 5: Stag Hunt Choices for a Randomly Chosen Agent of Each Choosing Type

From Figures 5a and 5b, we can evaluate that, in a static environment, the choosing Agents very quickly adapt to the near-optimal policy, voting for Hare when paired with the Hare-Brained Agent (explaining the remaining Hare votes after convergence), and Stag at virtually all other points. This demonstrates that the models are effectively tuned. For all Stag Hunts, regardless of Gossip score, the blue curves were virtually identical, with only slight variations at the very beginning, when the Agents had no gathered data to act upon. These graphs, despite being unremarkable alone, are useful as a precedent, showing that Agents can reliably settle on an optimal policy, given static and predictable conditions, regardless of the confounding effect that tuning to one-another's preferences causes.

4.4.2 Bison Hunt

Figure 6 demonstrate differing behaviour, not just between choosing Agent types, but also in behaviour as it pertains to adapting to a dynamic environment. The Agents are not quite converging to the ideal policy, possibly due to the rate of Epsilon decay, or possibly due to overtuning at the beginning of the simulations. Given that the more explorative Agents converged to a worse policy, it can be inferred that it is a result of the latter. The comparative slowness of the curve follows quite closely the rate of bison reward decay, as can be seen in the reward, although the rate of change in Bison values is much higher than that of the curves shown in these graphs. All of this is to show that, while Agents can reliably adapt to a dynamic environment, they can take a long time to do so, and the policy that they change to will likely be reasonably inferior to the optimal policy.

In the Holistic Evaluation section (Section 4.7), the ways in which this could have potentially been avoided will be looked at, as well as how these could have effected the nature of the other Hunts in the experiment.



(a) Average Agent Hunt choices over time

(b) Risky Agent Hunt choices over time

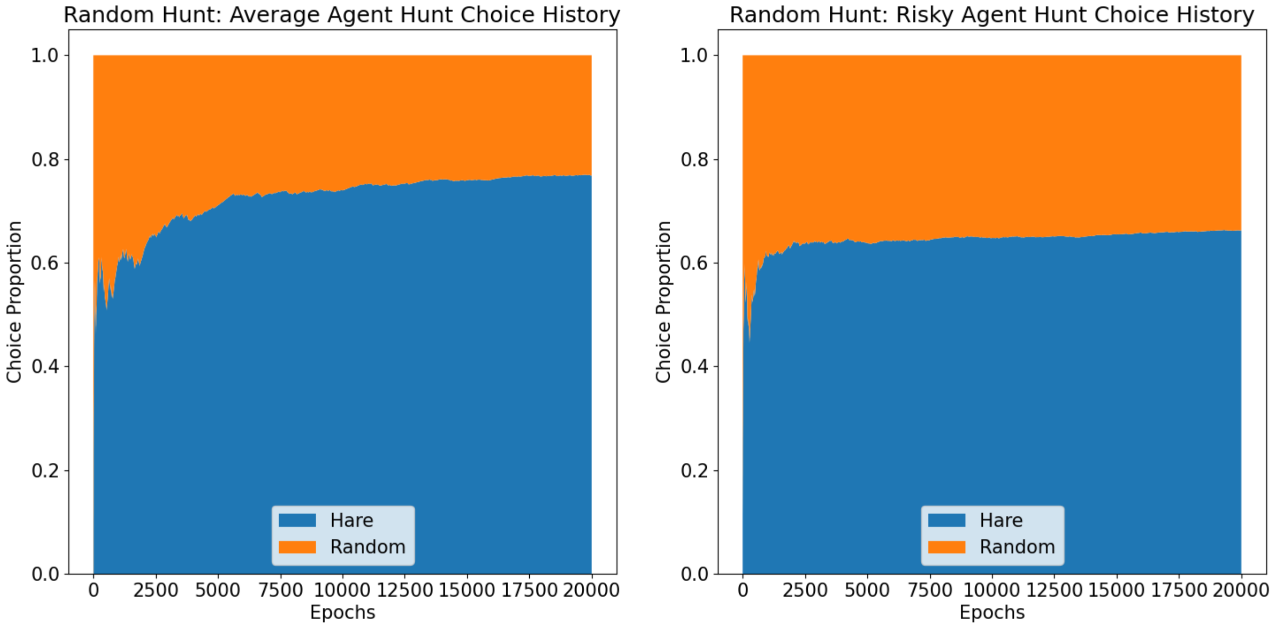
Figure 6: Bison Hunt Choices for a Randomly Chosen Agent of Each Choosing Type

4.4.3 Random Hunt

Figures 7a and 7b show something unique to the other two Hunt types, which is the convergence towards a suboptimal policy, even in the face of an environment wherein change is, while rapid, also consistent (pseudo-randomised between only 9 values). It can be inferred that this is a result of a feedback loop inherent to the Agents which stems from the inherent negativity bias and risk aversion of RL Agents in dynamic and unpredictable environments. The proposed feedback loop is as follows:

- Some Agents discover high rewards by voting Random, and some do not
- Those Agents who do not receive a comparatively high reward in the early stages turn to voting hare, whereas those that do continue to vote Random
- Those Agents who do turn to voting Hare pair with Agents who vote Random, leading to an advantageous outcome for the Hare voter and a disadvantageous outcome to the other
- The Random voter learns that this Agent is not safe to vote Random with in future, and therefore do not, leading to two Hare votes whenever the original Hare voter is present.

Risk aversion, then, can be viewed as a trait which is dominant in this system. It can be observed, however, that some Agents, primarily the Risky ones, vote Random frequently enough to have collated enough experience with Random outcomes to realise that it averages to a stronger outcome. Given that this is the case, the question of why the Agents have such a high margin of Random votes regardless, and that is likely because the same Agents will have both positive and negative experiences of voting Random that they relate to the Agent they were with at the time, meaning that the chance of a relatively consistent positive outcome voting Random with some Agents is, while low, still present and still lucrative.



(a) Average Agent Hunt choices over time

(b) Risky Agent Hunt choices over time

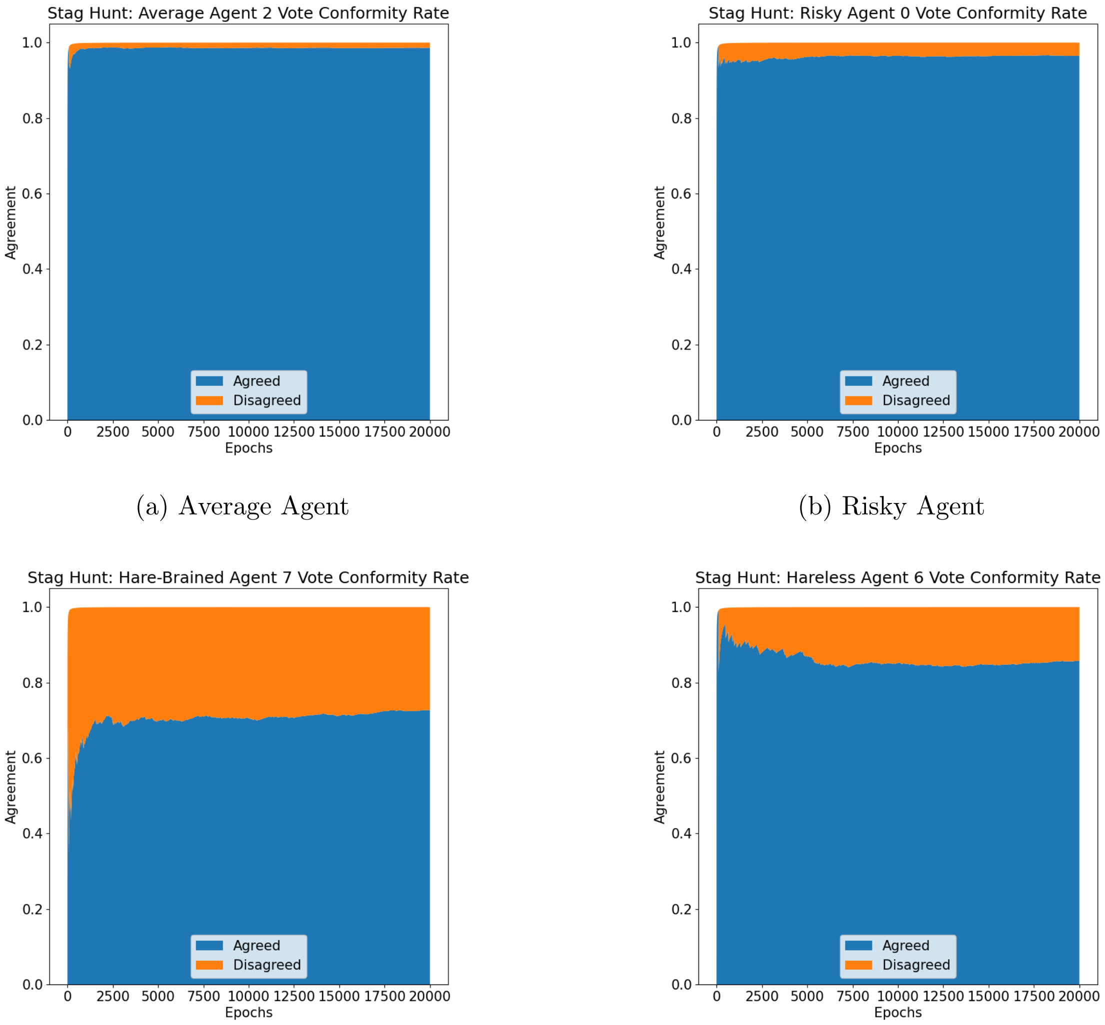
Figure 7: Random Hunt Choices for a Randomly Chosen Agent of Each Choosing Type

4.5 Rate of Vote Conformity per Agent Type

The graphs in this section will show the agreement rates of Agents over time, again in the form of a stackplot. This data, combined with the data in Section 4.4, will give a deeper insight into how Agents behaved in their interactions, and how well they adapted to interactions as a whole. It is worth noting that all four Agent types will be shown, since the difference in partner and difference in the nature of the choice mean that they may still afford interesting and novel insights. Since Hare-Brained and Hareless Agents will be included, it's worth noting the expected related agreement rates between the two. Since neither Agent will change their votes, they will, by default, always disagree when paired with one another. This will mean that the minimum expected disagreements will be, at any given epoch, 1/8 of all votes in all Hunt types. The other Agents, however, have no necessitated minimum disagreements, nor indeed minimum disagreements.

4.5.1 Stag Hunt

The graphs shown in Figure 8 show an interesting pattern in Hare-Brained and Hareless Agents, in that Hareless Agents converge to a higher than predictable proportion of disagreements, and the opposite is true for hare-brained Agents. This is consistent across reputation proportions. What this might imply is that the smaller the difference between potential outcomes, the more Agents struggle to converge on the optimal outcome. This, combined with the exploration rates giving the Average and Risky Agents more negative experiences, and, given the propensity for Hare-Brained Agents to pair more regularly with Risky Agents (as is elaborated on in the following section), their higher exploration rates lead to an increased rate of disagreement. If this is the case, the inverse can be assumed of the Hareless Agent, which is to say they benefit from a higher proportional rate of average Agent interaction, meaning they benefit from a higher level of predictability and training that is well suited to this static environment.



(a) Average Agent

(b) Risky Agent

(c) Hare-Brained Agent

(d) Hareless Agent

Figure 8: Vote Conformity for Stag Hunt

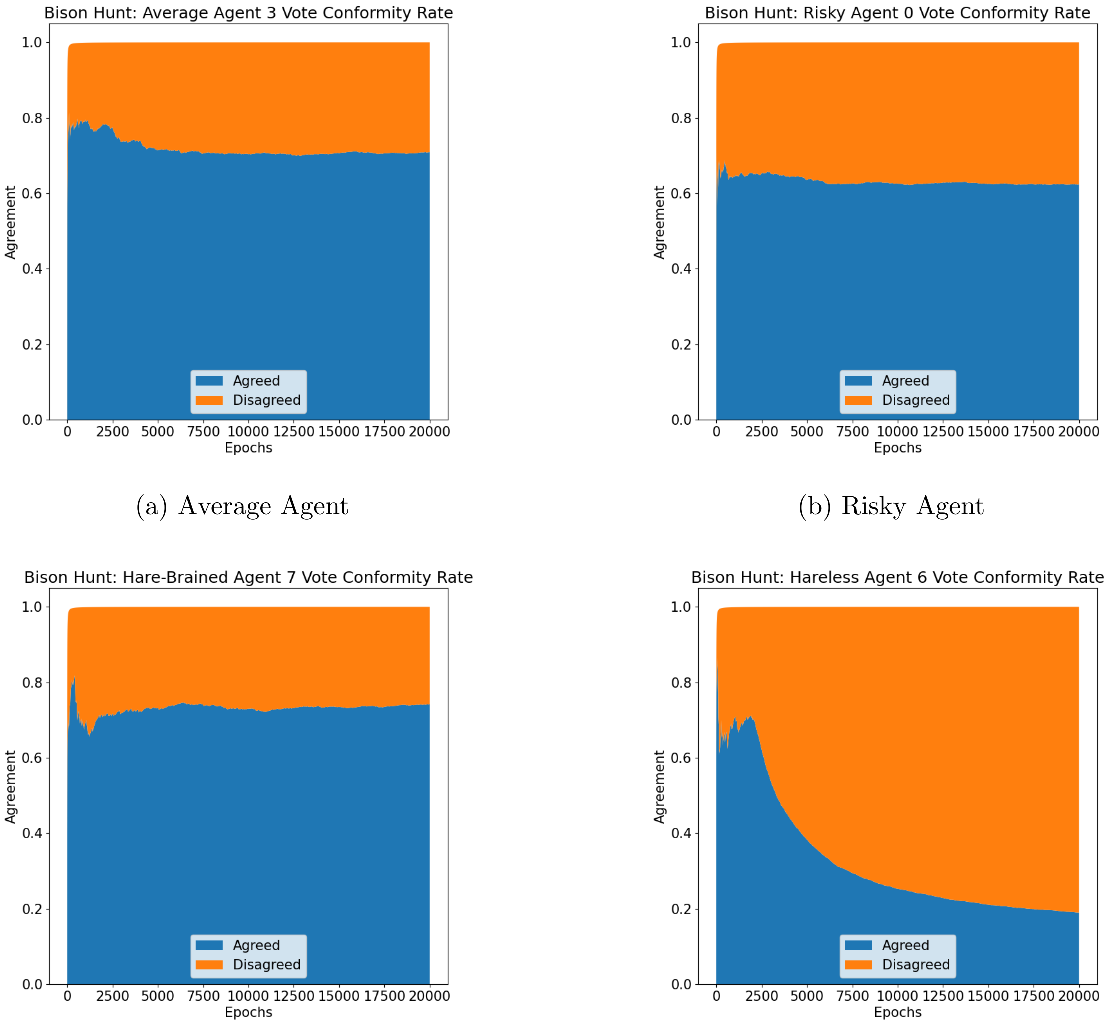
4.5.2 Bison Hunt

The graphs shown in Figure 9 shows similar patterns as are seen in Figure 8, in terms of imperfect convergence for the Hare-Brained Agent. This makes sense in this context, since there remains the obvious issue of matching with the Hareless Agent, but as well as this, the reward decay means that Average and Risky Agents may not be as well suited to find an optimal policy.

The rate of disagreement of the Hareless Agent has the similar rate of change as can be seen in Figure 6, which makes sense, and traces the rate of decay of the reward for Bison, again showing the ability of Agents to follow the change in environment effectively. Worth noting is that, in the Bison Hunt, disagreement is a measurement of success when pairing with the Hareless Agent, which explains the Average and Risky Agents' higher rate of disagreement.

4.5.3 Random Hunt

The graphs shown in Figure 10 show an interesting pattern in that Hare-Brained and Hareless Agents have a very similar, and relatively high level of agreement, accounting for their interactions with each other. This implies that the increase in disagreements seen in Average and Risky Agents are not with the Hare-brained or Hareless Agents but are in fact between one another. This is where the



(a) Average Agent

(b) Risky Agent

(c) Hare-Brained Agent

(d) Hareless Agent

Figure 9: Vote Conformity for Bison Hunt

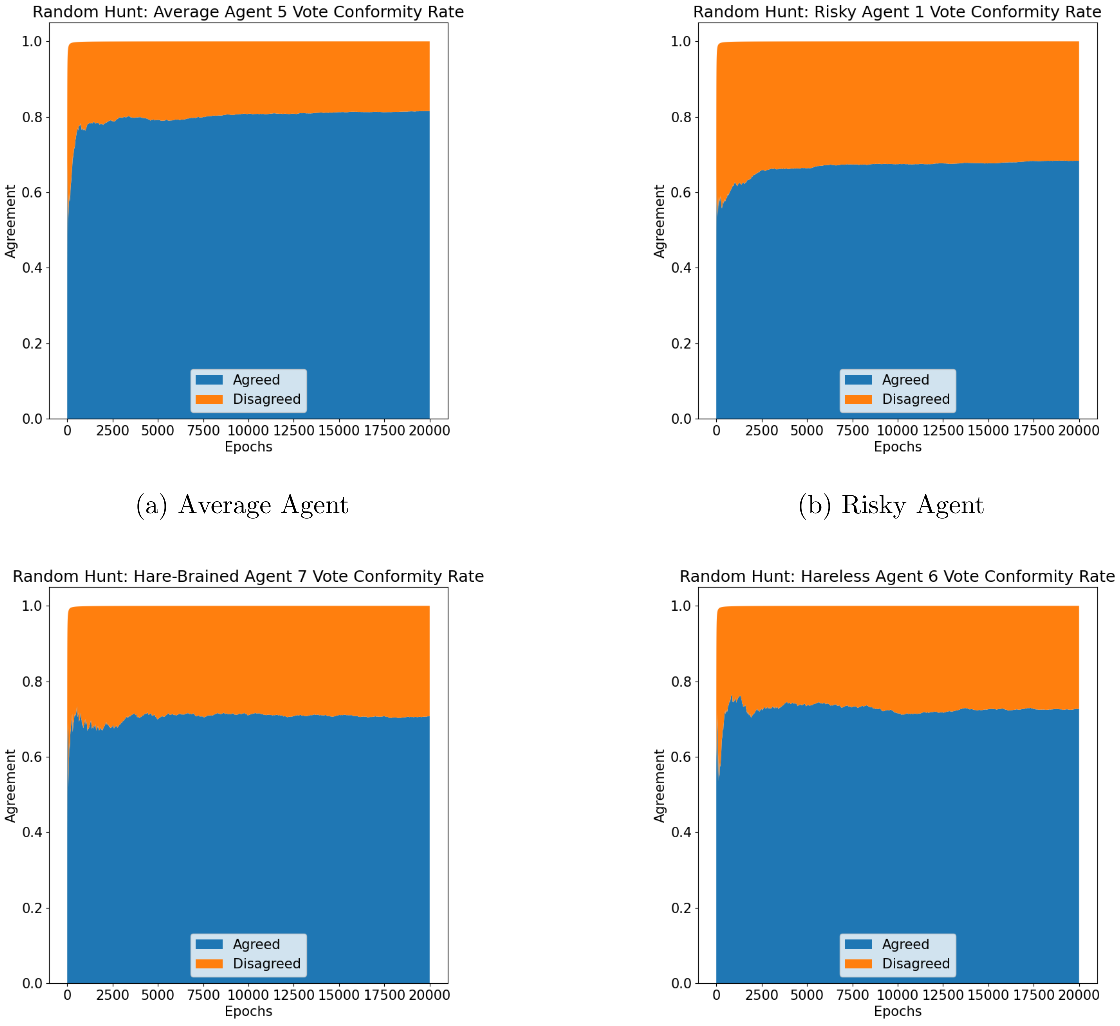
differences in exploration rate which previously made little difference lead to convergence to slightly different policies. This is consistent with the feedback loop outlined in section 4.4.3, and indeed, the disagreement rates are similar between Figures 10b and 7b to the choices made, making it quite clear that the Random votes are the typical point of contention.

4.6 Partner Choices

Choice of partner, being the primary way in which Agents differ behaviourally in this experiment, has two means by which it has been measured: Over Time per Trial and a Flat Average over all of them. These will be split into their own sections so as to separate the discussion by scale.

4.6.1 Over Time

The means by which reputation affects partner selection throughout the experiment remains somewhat unclear. This may be due to the randomness inherent to the behaviour of the Agents, the fluctuations in environmental feedback or both. Patterns are difficult to discern, though it would seem that often higher Gossip values correlate to a higher range between partners, as shown in Figure 11 (note that vertical scale is measures automatically, and is thus not uniform between graphs). This is a very



(a) Average Agent

(b) Risky Agent

(c) Hare-Brained Agent

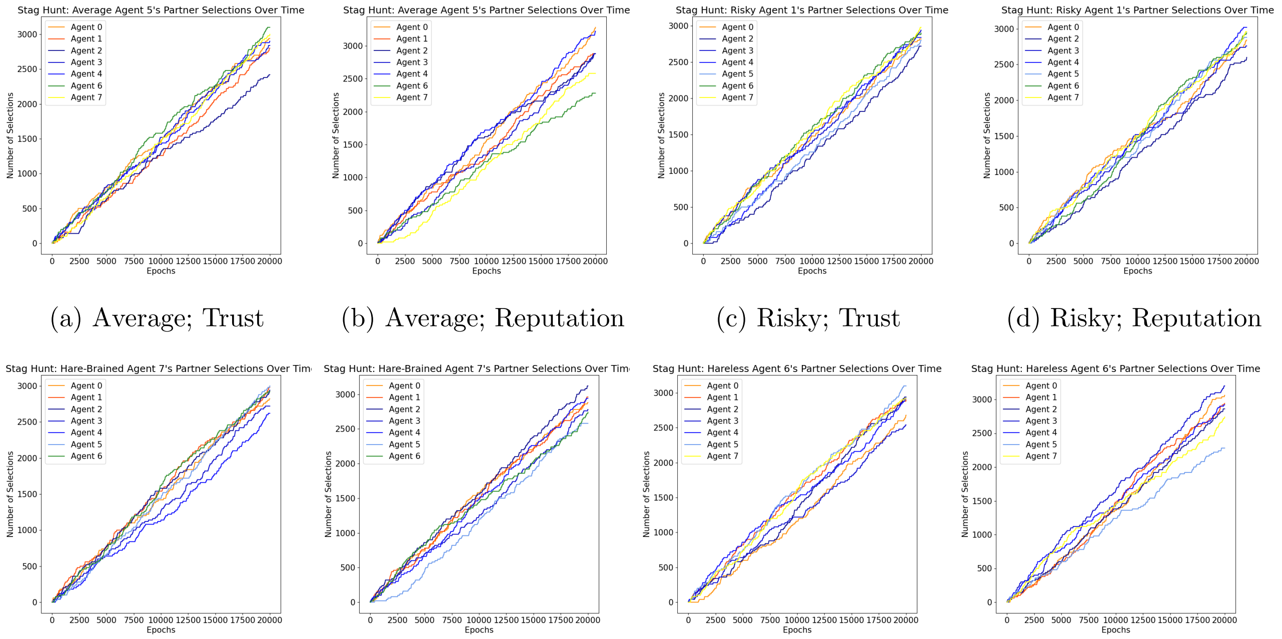
(d) Hareless Agent

Figure 10: Vote Conformity for Random Hunt

weak pattern, however, and this holds equally true for each Hunt type. This apparent chaos, is not necessarily a sign of a poorly adapted system, however, since learning the habits of each partner candidate is a vital part of keeping the system resilient, particularly when there are several ways that an Agent could be partnered with a suboptimal candidate.

4.6.2 Flat Averages

This measurement takes the average number of interactions with each possible partner the subject Agent experienced across the entire experiment, separated by Gossip Value and Hunt type. Since average and risky Agents all adapt behaviourally to one another, it is safe to consider their interactions with one another in any given trial interchangeable with any other Agent of the same type. As a result, the table shown in this section looks at the Hare-brained and Hareless only, to better understand how they adapt to each environment, and how it adapts to them. Looking at both Table 1 and Table 2, it can be observed that, among the lower Gossip levels, the two Agents regularly pair up with one another most often, which is abjectly poor in terms of overall gain, and increased reputation seems to mitigate this fact somewhat. This may be a worthwhile area for further exploration, since using reputation as a metric to encourage even-handed exploration of other partners may enable Agents to



(a) Average; Trust (b) Average; Reputation (c) Risky; Trust (d) Risky; Reputation

(e) H-B; Trust (f) H-B; Reputation (g) Hareless; Trust (h) Hareless; Reputation

Figure 11: Stag Hunt: Partner Choice History per Agent Type, with Trust Only Vs Reputation Only

find superior partners without otherwise affecting their ability to make decisions in the future, as one necessarily must in order to implement Epsilon Decay. Another observation that can be made is how, around Gossip Value 2, the range between Agent frequency becomes much smaller, While this could be a coincidence, the correlation between the two tables could lead one to imply that being a risky Agent benefits from these middling Gossip scores in some way. This, however, is primarily speculative, as the output is very noisy.

	Agent 0	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5	Agent 7	Range
0	2852	2810	2856	2754	2870	2796	3062	308
1	2720	2814	2880	2858	2872	2690	3166	476
2	2786	2866	2786	2810	2914	2886	2952	166
5	2806	2932	2876	2760	2902	2806	2918	172
10	2908	2950	2780	2942	2806	2800	2814	170
100	2906	2892	2826	2838	2848	2864	2826	80
Inf	2892	2962	2862	2888	2800	2734	2806	228

Table 1: Stag Hunt: Hare-Brained Agent 6 Average Partner Frequency Per Gossip Score

4.7 Holistic Evaluation

The results of this section show that both systems of partner selection that these Agents utilise, while imperfect, are very effectively adapted to static, or slowly changing environments, but struggles when exposed to a rapidly oscillating dynamic environment, which leads to convergence on a highly suboptimal policy. One pattern that can be observed within these results are that the Risky Agents were more easily swayed towards suboptimal policies in dynamic environment, suggesting that an exploration rate of 0.3 is unsuited to these environments. This is unsurprising, and a matter of tuning, and as such, is something of an asset to this experiment. It provides Agents whose reputation

	Agent 0	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5	Agent 6	Range
0	2940	2826	2810	2776	2752	2834	3062	310
1	2852	2840	2738	2820	2814	2770	3166	428
2	2756	2990	2772	2772	2892	2866	2952	234
5	2844	2866	2860	2914	2790	2808	2918	128
10	2918	2826	2804	2848	2896	2894	2814	114
100	2784	2968	2864	2858	2860	2840	2826	184
Inf	2860	2828	2962	2912	2898	2734	2806	228

Table 2: Stag Hunt: Hareless Agent 7 Average Partner Frequency Per Gossip Score

is likely to be inferior, encouraging repeated pairing towards the end of each trial if in Bison Hunt or Random Hunt. This can be observed in Table 2, wherein the two Risky partners are seen to be have near-maximum or maximum selection frequency, creating a trend wherein as reputation increases, the more Risky Agents tend to pair with one another, or the worst of the Static Agents (Hare-Brained or Hareless) for the given Hunt type at that time.

5 Conclusions and Critical Evaluation

The stated purpose of this experiment was to observe and measure the distinctions and mechanical implications of Trust and Reputation. While, in some ways this has been achieved, it could be considered insufficient in others. While the manners by which Agent behaviour changes given different proportional consideration of reputation are broadly negligible, this is not a meaningless discovery, however, since there are other secondary advantages to using a Reputation system rather than a Trust-based one.

The primary advantage for a Reputation system, then, might be in matters of memory and performance, it being cheaper to store each Agent's score in one place and have each epoch update every Agent each time than to have each Agent store their own data. While in this experiment, the Reputation system served only to distribute Agent pairings slightly more evenly, it's possible that, given a task with a higher decision space and more meaningful reasons for particular Agents to interact with one another, a reputation system may be used to better isolate less fit Agents and allow for more fit matchmaking. Regardless, the fact that, armed with the Reputation system, Agents were able to converge to such effective policies shows that the system has potential as a matchmaking system for RL Agents, which may speed up more complex systems by a significant amount.

A notable drawback to this experiment lies in the nature of the problem the RL Agents were intended to solve. The minimal decision space provided by the Stag Hunt and its derivatives have meant that the task may have been too trivial to effectively test the distinctions between Trust and Reputation systems. It's possible that this may be due to an over-reliance on direct reciprocity as an analogue for trust-building (Isoni and Sugden 2019).

Another potential drawback is the similarity between Agents. It is possible that Agents need to be more distinct from one another (i.e. in matters of discount factor or learning rate) in order to give further metrics by which some less fit Agents will be ostracized via poor reputation rather than have their maladaptive policies proliferate through the group. It might also be pertinent to have Agents' interactions with some candidates affect their decision-making with other, unrelated Agents in some way. It is also true that most Agents are able ton adapt to their environments more or less regardless of their chosen partner, which is a problem when sensible long-term partner choice is what is to be examined. The scope of results is also not particularly fit-for-purpose, where it comes to discussion. Data gathered is predominantly either very weakly correlated or resilient to the proposed experimental changes.

Nonetheless, this experiment has shown, if nothing else, the viability of Reputation as a means of disseminating partner information for use in matchmaking RL Agents in cooperative systems, and that it is able to compete with the more isolated Q-learning process typical to RL models.

5.1 Further Development

The research presented in this report could be developed by having similar Agents navigate a more strategic game wherein more distinct emergent characteristics may be examined, and where the next state of each Agent can be controlled by said Agent, rather than being controlled entirely by an intermediary mechanism (in this case returning to partner selection after each round of Hunts). A cooperative game that fits this description may require that three or more Agents have to form a group and each Agent must pick one of three roles, and only when each Agent chooses a unique role from the other two are they rewarded, though the matter of complexity could scale much further than is covered in this report.

References

Ahlstrom, Laura. 2023. “Stag Hunt.” *Inomics* (July). https://inomics.com/terms/stag-hunt-1537413.

Antony, Snobin, Raghi Roy, and Y Bi. 2023. “Q-Learning: Solutions for Grid World Problem with Forward and Backward Reward Propagrations” [in English]. In *AI-2023 Forty-third SGAI International Conference on Artificial Intelligence CAMBRIDGE, ENGLAND 12-14 DECEMBER 2023.* http://www.bcs-sgai.org/ai2023/.

Barclay, Pat. 2015. “Reputation.” Chap. 33 in *The Handbook of Evolutionary Psychology,* 1–19. John Wiley & Sons, Ltd. ISBN: 9781119125563. https://doi.org/https://doi.org/10.1002/9781119125563.evpsych233.

Bowyer, Caleb M. 2022. “Strategies for Decaying Epsilon in Epsilon-Greedy.” *Medium,* https://medium.com/bankless-dao/what-is-a-trustless-system-3ded568c8921.

Chen, James. 2024. “Nash Equilibrium: How It Works in Game Theory, Examples, Plus Prisoner’s Dilemma.” Edited by Gordon Scott and Kirsten Rohrs Schmitt. *Investopedia* (June). https://www.investopedia.com/terms/n/nash-equilibrium.asp#toc-prisoners-dilemma.

Isoni, Andrea, and Robert Sugden. 2019. “Reciprocity and the Paradox of Trust in psychological game theory.” *Journal of Economic Behavior & Organization* 167:219–227. ISSN: 0167-2681. https://doi.org/10.1016/j.jebo.2018.04.015.

Karunakaran, Dhanoop. 2020. “Q-learning: a value-based reinforcement learning algorithm.” *Medium,* https://medium.com/intro-to-artificial-intelligence/q-learning-a-value-based-reinforcement-learning-algorithm-272706d835cf.

Kohler, Tanner. 2023. “10 Survey Challenges and How to Avoid Them.” *NN Group* (February). https://www.nngroup.com/articles/10-survey-challenges/.

Korsgaard, M Audrey. 2018. “Reciprocal trust: A self-reinforcing dynamic process.” In *The Routledge companion to trust,* 14–28. Routledge.

Liu, Yuhong, and Yan Lindsay Sun. 2014. “Securing Digital Reputation in Online Social Media [Applications Corner].” *IEEE Signal Processing Magazine* 31 (1): 149–155. https://doi.org/10.1109/MSP.2013.2282414.

Pan, Xiaofei Sophia, and Daniel Houser. 2013. “Cooperation during cultural group formation promotes trust towards members of out-groups.” *Proceedings of the Royal Society B: Biological Sciences* 280 (1762): 20130606. https://doi.org/10.1098/rspb.2013.0606.

Resnick, Paul, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. “Reputation systems.” *Communications of the ACM* 43 (12): 45–48.

Ruhl, Charlotte. 2023. “Cognitive Bias: How We Are Wired to Misjudge.” Edited by Saul McLeod and Olivia Guy-Evans. *Simply Psychology* (October). https://www.simplypsychology.org/cognitive-bias.html.