The role of reward, risk, and behavior in trust formation and collaboration among multi-agent systems

---

A thesis presented to

the faculty of the School of Engineering and Applied Science

The University of Virginia

In partial fulfillment

of the requirements for the degree

Master of Science, Systems and Information Engineering

Nikesh D. Kapadia

ndk5bd@virginia.edu

Advisory Committee:

Peter A. Beling (Chair)

Matthew S. Gerber (Advisor)

Charles A. Holt

Stephen C. Adams

April 2019

Abstract

The concept of trust is diverse and widely used to understand dynamics within multi-agent systems (MAS). Various academic disciplines study trust to understand the interactions and decisions of humans and/or artificial agents. We define trust as the extent to which an agent is willing to take on the risk governed by the behavior of another agent.

The following research formulates trust as a decision process under the reinforcement learning (RL) framework. Distinct from previous work, trust is formalized as an action enabling meaningful measurement of the construct as the expected return with consideration to the variance of the partner's behavior. The framework facilitates the investigation of the role of reward, risk, and partner behavior within trust formation and collaboration between the agents. We examine these characteristics among two agents operating in a gridworld simulated environment.

We find that having information on the partner's behavior, and the ability to take risks are crucial aspects for trust formation. When agents make risk-conscious decisions upwards to 62.54 % rates of mutual collaboration can be achieved. However, there is a trade-off where high values of trust can lead to over-trust situations; situations where one agent trusts the other agent to its own detriment. Then, the agent must adapt how much risk it is willing to assume, to control for these mis-coordinated outcomes.

We propose several avenues for future work in which the framework estimates and integrates risk into the agent's decision-making process. The framework can be used to further articulate interdependencies and the characterization of interactions, and expanded to larger multi-agent systems.

Contents

1 Introduction

Multi-agent systems are a way of describing a collection of autonomous agents that each have their own goals, missions, or responsibilities [1]. These systems are growing to become pervasive in society. We are using increasingly intelligent robots and intelligent tools to interact with each other and interact with people, with everyone having their own goals. Natural questions that come up are how can we get everyone to synchronize efforts, and how can everyone work together towards an even greater mission?

In the domain of behavior change and intervention systems, machines are in a relationship with their humans. For example, an intelligent artificial pancreas encourages beneficial health behavioral changes to the human patient [2]. Does the human listen? I trust my machine, but I trust my other instincts more. When would the human trust the machine's recommendations? In an education setting, a intelligence robot is providing instruction to a classroom [3]. Each student will inculcate the instruction so far as how much he or she trusts the teacher.

Larger and distributed systems comprise of a complex network of relationships. Energy is being harvested from a distributed network of sensors [4]. Individual producers must coordinate with each other to pool the commodity and coordinate with buyers. In the manufacturing realm, robots and human operators perform specialized tasks [5]. Whether it is a smart energy market or a factory setting, individual agents cannot blindly trust each other. However, trust is required for collaboration to a greater success. What are the risks that each agent has to accept when they partner, and how are those risk mitigated? How are these teams built?

Multi-agent systems proliferate in society; their capabilities and their associated challenges. Collaboration between autonomous agents and humans is significant challenge. My thesis is, **Trust is the primary determinant of collaborative outcomes within multi-agent systems.**

2 Related Research

The concept of trust is diverse and widely used to understand dynamics within multi-agent systems. Various disciplines from economics, psychology, sociology, and computer science study trust to understand the interactions and decisions of humans and/or artificial agents. Cognitive factors such as fear and hope can impact the construct of trust, as well as social and environmental

conditions. Trust is dynamic and changes over time between agents, and depends on specific situations or circumstances. These characteristics make trust a challenging construct to model yet vital to understand social and network interactions. Among these factors, measuring trust becomes significant as well.

Various academic disciplines study trust and develop definitions to characterize the construct. Organizational behaviorists Rousseau et al. developed a cross-disciplinary definition to promote shared understanding across all academic domains. Rousseau et al. define trust as, "a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of an other" [6]. Rousseau and other scholars note that trust is comprised of a psychological and social component [6]–[8]. A psychological process is present that drives an individual to a decision based on contextual information and emotions. A social process is required where the individual must consider to interact, collaborate, or dependent on a partner. Mayer et al. define trust as, "the willingness of a party to be vulnerable to the actions of another party" [9]. The sociologist Niklas Luhmann described trust as the willingness to take risk under uncertainty [10]. Scholars note the central role that consideration of risk or reliability of the partner in decisions of trust [6], [11].

A review across multiple disciplines in regards to trust related research illustrate the key characteristics of trust include interdependence, risk, a psychological process, and a social interaction. How do we integrate these elements into a framework to model trust formation and collaboration between multi-agent systems (MAS)? The following subsections elaborate on current approaches to study these characteristics, how current computational approaches model trust within multi-agent systems (MAS), and their limitations to study these characteristics.

2.1 Firefighter Scenario

Before I talk about those trust characteristics, I want to introduce a simple example to reference through out the presentation. The scenario is illustrated in 1. Bob is in his office building, which is on fire. A firefighter contacts him and says, "The bottom floor is on fire. We have to exit out of the roof. Meet me on the roof, I'm going to check the rest of the building for other people." Figure 2 articulates Bob's thought process. Bob has a choice, should he trust the firefighter and go to the roof. What if the firefighter does not show up, he could die. Or, Bob can say, "I don't

Figure 1: The firefighter scenario

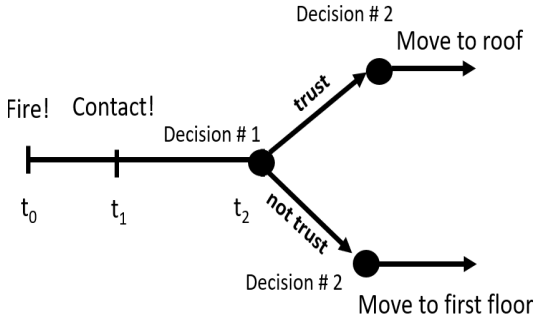

Figure 2: Bob's decision timeline

believe the firefighter. I know my building better than him. I can find a way out of the first floor myself." How does Bob negotiate this dilemma? Should Bob trust the firefighter? Well it depends, but what does it depend on ?

2.2 Socio-Cognitive Approach

The significant challenge was to map important characteristics of trust determined from psychology and sociology, into an integrated computational framework. Circa 1995, Cristiano Castelfranchi and Rino Falcone introduced a socio-cognitive approach to model trust in multi-agent systems [7]. Most contemporary computational trust models lean on the work of Castelfranchi and Falcone. They integrate four primary characteristics, their framework to describe trust development.

First, the agent has its own goal seeking behavior. Second, The agent forms a prediction or belief on the partner's behavior. Third, the agent makes a decision or intention to be vulnerable to another's actions. Fourth, the agent induces a subjectivity created through a cognitive process. Factors of fear, hope, beliefs, and attitudes impact the agents perception of the reward, and willingness to take risk.

How do these characteristics map into the firefighter scenario in figure 1. First, Bob is motivated by is own interests. He is going to act in the manner to save his own life. Second, there is a component of partner prediction. Bob needs to predict the firefighter's behavior. What is the likelihood that the firefighter will meet me on the roof? Third, Bob has to make a decision, go up and be subject to what the firefighter does, or go down and be irrelevant to what the firefighter does. Fourth, there is an induced subjectivity. Bob's own inherent willingness to take risk, and his own emotions effect his decision-making process.

The work of Castelfranchi and Falcone also demonstrate how the environment and context can influence an agent's decision-making through changes in the information, intrinsic beliefs, and/or observed beliefs. Finally, the authors make a distinction between trust and collaboration. Trust does not guarantee collaboration and vice versa. These implications illustrate how trust is a unidirectional, subjective, and dynamic process—important characteristics to understand trust development in multi-agent systems (MAS). The research paved the way for follow on work on how trust is formed through specific information sources and in specific MAS application areas. The limitation of the framework will be discussed at the end of the next subsection.

2.3 Computational trust models for MAS

Within the last ten years, there has been an extensive amount development in computational trust models for MAS. Many models informed through the work of Castelfranchi and Falcone. Each model considers a different approach in calculating trust among interactions in a variety of application areas. In large multi-agent networks, the key task for an agent is to often find a trustworthy agent to help advance towards a goal. Trust models will assist with that task and primarily use quality of direct interactions as a way to evaluate and calculate trust. When direct interaction information is not existent or is insufficient, trust models utilize third party or witness information to assess the trustworthiness of an agent [12]. Sabater and Sierra developed the foundational ReGreT model [13]

demonstrating how trust can be calculated in a large social network using four pieces of information: direct interactions, third-party information, social relationships, and system reputation [14]. Other models may utilize additional sources of information such as certification protocols [15], update trust in a probabilistic manner [16], or a Bayesian network approach [17]. Overall these approaches consider multiple sources of information into trust formulation.

The framework posed by Castelfranchi and Falcone and the above computational models focus on the calculation of a trustworthiness score. They do not facilitate the study of interactions, trust decisions, and associated trade-offs. There is a key characteristic missing from these approaches, the characterization of interdependence.

2.4 Interdependence Theory

Interdependence is a situation when the goals or interests of two agents overlap, and cannot be achieved without the agents relying on each other [11]. Psychologists Harold Kelley and John Thibaut first introduced interdependence theory in 1959. The theory formalizes interdependent rewards and interpersonal interactions within contextual (environmental) conditions [8], [18]. The theory allows for the study of interdependence along multiple dimensions such as degrees of dependence, influence, and dependability [8]. The theory explains how interdependent rewards can be characterized as a cost and a benefit, which are dictated by the environment or circumstance. The two researchers introduced game theoretic matrices to articulate the interplay between environment, decisions, and outcomes. The matrix approach illustrates the utility of studying trust as interactions, however it is limited to single interaction games. The theory provides an stepping point to settings that involve sequential decision making, and settings that involve more than two agents.

In the firefighter example in figure 1, Bob must rely on the firefighter, to a certain degree. Interdependence theory helps us characterize this dependence. When Bob needs to make a decision to be vulnerable, he compares the outcome from being dependent on the firefighter, with the outcome of acting independently (alone) by going downstairs.

Wagner et al. studied trust development between humans and machines through stochastic game theoretic models. The team experimentally tested Kelley and Thibaut's interdependence theory in the human-machine interaction domain. Wagner et al. examined the extent to which interdepen-

dent reward structures and inherent risk disposition influenced trust decisions [19]. Gaps between theoretic predictions and experimental results were attributed to human cognitive responses to different situations. Individuals react differently (emotionally) than each other in different situations.

Game-theoretic approaches to study trust development can be improved in three ways to include learning, delayed feedback, and dynamic situations. Game-theoretic approaches do not adequately address trust formation as learned behavior between interactions. The analysis of the development of trust through a learning model may yield additional insight on how and why agents deviate from theoretic outcomes. Do situations and previous interactions shape trust development in a way that ultimately leads to sub-optimal decision-making? Maybe Bob has a history of being betrayed by firefighters, which leads him to not trust this one.

Delayed feedback will also improve the framework. Many times, the feedback from the decision to trust is not received until later on. Bob does not immediately know that he made the right decision to trust the firefighter. He does not get the reward until the end of this scenario. Finally, dynamic situations will help improve the characterization of trust. Trust is dependent on the situation. Bob's trust in the firefighter is different if he is closer to the roof vs closer to the first floor, as well as other details that may characterize this environment.

2.5 Markov Decision Processes

Trust models built within Markov Decision Processes (MDPs) characterize trust as a learned process, include delayed feedback, and include dynamic situations. Chen et al. capture the relationship through a partially observable Markov decision process (POMDP) [20]. Within the model, trust is defined as the approximation of the history of interaction between two agents. Therefore, the probability of a particular action is conditioned on the state and the trust parameter, which is updated after each interaction. The agent will tend to select particular actions depending on the level of the trust parameter.

Reinforcement learning (RL) is a form of machine learning characterized by goal-oriented behavior within a Markov decision process [21]. RL assumes the agent's behavior is driven to maximize reward. The assumption remains consistent with previously specified forms of trust research where people are often modeled as reward maximizers [8], [7]. The RL framework allows for stochastic decision-making and feedback from the environment, which characterizes the uncertainty condi-

tions of rewards. Decisions based on contextual information and decisions based on interactions are effectively tested in RL. Therefore, trust conditions such as risk, reward, and interaction can be mapped into the RL framework. Furthermore, the use of a learning model provides insight on trust formation as through repeated interactions. Trust is maintained as a parameter that is learned through direct interactions [22], [23], and witness information [24], [25]. In large scale MAS, agents are classified based on their trustworthiness score to prioritize interactions. Classification mechanisms are executed through heuristics [22] or fuzzy logic systems [24].

The above trust models focus on the four trust characteristics posed by Castelfranchi and Falcone. However, there is no characterization of interdependence- how much does the agent have to depend on the partner versus not have to depend on the partner. The lack of this component limits our ability to study risk in the context of trust relationships.

Moreover, the above trust models formulate trust as a parameter. There is a disconnect between actual and accepted definitions of trust out of psychology and sociology, and how the definitions are formalized in these models. Trust as a parameter does not meaningfully express the decision to be vulnerable or a willingness to take risk.

2.6 Safe Reinforcement Learning

There is an important relationship between the trust and risk. In traditional approaches to RL, agents only consider the expected return of outcomes. In trust situations, it is significant to consider the risk associated with the interaction. Approaches in safe RL balance maximization of expected return with risk, in order to motivate the agent to avoid high risk states despite their potentially high return. Safe RL approaches modify the optimization criteria by including the consideration of reward variance [26]. Heger introduced a minimax criteria where the agent maximizes the expected return over the trajectory of least variance [27]. On the other hand, the optimization criteria can be a linear combination of expected return and variance [28], [29], or an exponential expected utility function [30]. These changes in the optimization criteria provide a utility on the variance for the agent to consider at a given state. Other safe RL approaches modify the exploration process [26]. Apprenticeship learning or initial learning can guide agents through the learning process to avoid high-risk states. Agents can also be guided through the exploration process through a risk metric identified for each state-action pair [31]. Safe RL approaches primarily focus on minimizing risk in

control applications. The literature does not address encouraging agents to take risks to promote trust development and collaboration.

2.7 Expected Utility Theory

Several of the safe RL approaches above are informed by expected utility theory. These approaches acknowledge that an agent's perceived value of a reward, and the real value of a reward are not necessarily the same. The gap between utility and real value is due to a inherent subjectivity of the agent. The subjectivity can be influenced by a variety of cognitive emotions such as fear, hope, regret, or caution [7], [32]. Expected utility theory calls this subjectivity parameter a risk preference that controls the shape of the utility function. Safe RL approaches have directly used expected utility functions [30] in the objective criteria. Also, several safe RL approaches contain a parameter to control the amount of variance to consider as part of the reward expectation calculation [28], [29]. The approaches of expected utility theory integrated into the RL architecture are focused on risk aversion and safety. In trust interactions, we seek to encourage risk seeking behavior to promote trust development.

3 Theoretical Formulation

Various academic disciplines have conducted important research into trust. A survey of the literature indicates that key elements of trust are interdependence, risk, a psychological process, a social interaction, and situational context. In order to study trust formation and collaboration among MAS, a framework must integrate these trust characteristics. The computational trust framework by Castelfranchi and Falcone best integrate these characteristics highlighting that trust comprises of a prediction of the partner, decision to be vulnerable to another, cognitive process, and goal seeking behavior [7]. However, the framework does not characterize interdependency, which limits our understanding of the nature of the interaction between the agents. Wagner et al. use a game-theoretic framework built on Interdependency Theory by Kelley and Thibault to better characterize interactions [18], [19]. However, both approaches do not consider trust development through learning, delayed feedback, and dynamic situations. Trust models build as MDPs address these issues, but formalize trust as a parameter. Trust as a parameter does not meaningfully ex-

press the decision to be vulnerable or a willingness to take risk. Therefore, we seek to develop a theoretical formulation that characterizes the significant trust elements, to allow us to study the formation of trust and collaboration among multiple agents.

3.1 Proposed Definition

We formalize our definition of trust as follows: **trust is the extent to which an agent is willing to take on the risk governed by the behavior of another agent.** Risk is defined as the variance in return [33], governed by the behavior of another agent.

If agent 1 trusts agent 2 a great deal, then agent 1 is willing to take on risk that is governed by agent 2's actions and will behave in accordance with the expectation that agent 2 will act in a particular way. In trusting agent 2, agent 1 makes its future rewards contingent upon the behavior of agent 2. On the other hand, if agent 1 does not trust agent 2, then it will do its best to eliminate agent 2's influence over agent 1's future rewards. Misplaced trust has negative utility when the agent 1 acts in a way that is sub-optimal given expectations of agent 2 are not met. The definition emphasizes the key role that risk and reward play in trust. Furthermore, the definition poses trust as a decision to take on risk with another agent, highlighting the interdependent reward and trust as a decision.

3.2 Problem Formulation

The proposed definition helps create a theoretical formulation as a Markov Decision Process (MDP). We build off of the firefighter scenario in figure 1. We formulate a two-agent system, $n = 2$, with Bob and the firefighter. Each agent has a discrete set of states in state space, $s_i \in S$, representing the location or floor of the agent. There is a discrete set of two actions available to the agent, to either move up or to move down, $a_i = \{a_U, a_D\} \in A$. There is an unknown probability transition matrix, T. Each agent is given its own reward function, R. There is a time-valued discount parameter, γ. The Markov Decision Process (MPD), $< n, S, A, T, R, \gamma >$ is created. Each agent follows a default ϵ-greedy policy, π, to balance exploration and exploitation. Each agent's objective is to maximize their own return (long term reward), which results in the optimal policy for the agent, π^*. Equation 1 is the objective function, where t is the time-step.

$$\pi^*(a|s) = \max_{\pi \in \Pi} \mathbf{E}_\pi[G_t|s_t, a_t] = \max_{\pi \in \Pi} \mathbf{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|s_t, a_t] \tag{1}$$

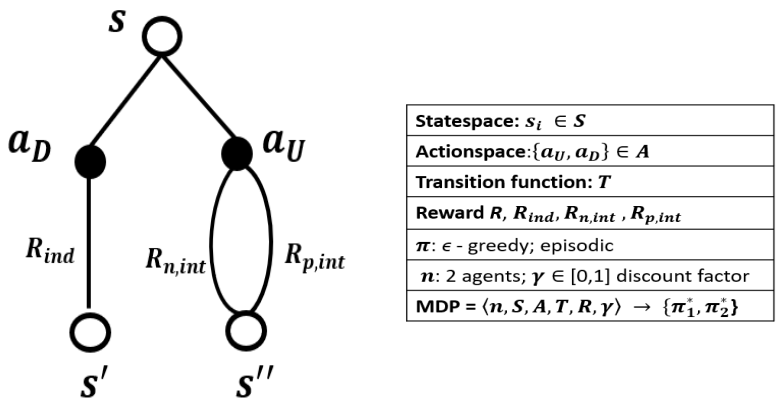

Figure 3: Simple problem formulation

There are two outcomes for a particular agent. Figure 3 visually depicts the agent's decision. Both agents hold the same perspective. From any state, the agent faces a choice. Let's look at the choice from the perspective of Bob. Bob can choose to go down (action down, a_D, to act independently, which will lead to a consistent reward noted as the independent reward, R_{ind}. Or, Bob can choose to go up (action up), a_U, which will lead to two possible rewards that are contingent on the other agent (the firefighter's) behavior, an $R_{n,int}$ or $R_{p,int}$.

These reward contingencies are noted as interdependent rewards. A reward contingency is defined as a set of conditions that both agents must meet in order to receive the positive interdependent reward, $R_{p,int}$. If both agents satisfy the conditions, it would be considered a successful interaction and both agents will receive the positive interdependent reward, $R_{p,int}$. If an agent deviates from one of the conditions, it will result in a failed interaction and the agent who attempted to collaborate will receive a negative interdependent reward, $R_{n,int}$.

If Bob goes up and the firefighter is already on the roof, that is considered a successful interaction and Bob will receive the positive interdependent reward, $R_{p,int}$.If Bob goes up and the firefighter is not on the roof, that is considered a failed interaction and Bob will receive the negative interdependent reward, $R_{n,int}$. Figure 4 articulates the reward structure in game-theoretic matrix form. In a 2x2 matrix, the interdependency between both agents is highlighted. The matrix also

highlights how this can be seen as a simple coordination game.

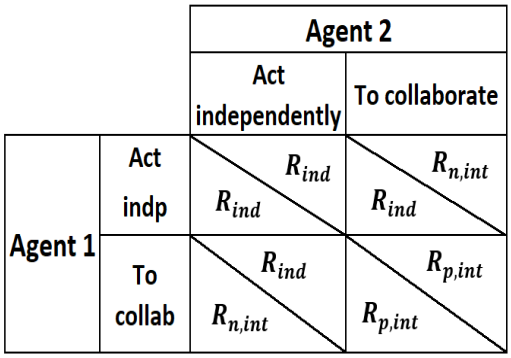

Figure 4: Theoretical Reward Matrix

Each agent must weigh the expected return of pursuing the independent reward, choosing a_D, and the expected return of pursuing the interdependent reward, a_U. By definition, the action-value function $q(s, a)$, is the expected return G_t, starting from state s, taking action a, and then following policy π [21].

$$q_\pi(s, a_D) = R_{ind} + \gamma \mathbf{E}_\pi[(G_{t+1}|s_{t+1}, a_{t+1})] \tag{2}$$

$$q_\pi(s, a_U) = \begin{cases} R_{n,int} & , if \quad C = NM \\ R_{p,int} & , if \quad C = M \end{cases} + \gamma \mathbf{E}_\pi[(G_{t+1}|s_{t+1}, a_{t+1})] \tag{3}$$

In equation 3, $C = NM$ denotes if the reward contingency conditions are NOT MET by both agents, versus $C = M$ denotes if the reward contingency conditions are MET by both agents. The condition in the firefighter scenario is if the partner is present at the terminal point. As equations 2 and 3 specify, the q-values comprise of an immediate reward and discounted future return. The expected return from action up, is the expected return from a policy pursuing the independent goal. The expected return from action down is dependent on the immediate outcome of the contingency, which dependent on the behavior of both agents. Because the expected return from action up is dependent on the behavior of both agents, there is variability on the expected return from the interdependent goal. On one hand, an agent can choose an action that eliminates any dependency on the other agent and can choose to pursue the independent reward, which has no variance due to the partner's behavior.

Standard RL methods in incrementally estimating the expectation do not adequately characterize the distribution of the interdependent reward. The expectation will be adjusted heavily based on the latest outcome, and not consider the history of outcomes as an indication of the other agent's behavior. The agent needs more insight on the partner's behavior in order to determine if pursuing the interdependent reward is a worthwhile endeavor.

There are two problems in the standard RL formulation. First, how do we restructure the action space to include the concept of trust as a primary construct? Second, how does the agent consider the variance (in outcomes due to the partner's behavior) in its decision-making process?

3.3 Trust as an action

Motivated by the proposed definition of trust, as well as trust scholars such as Meyer, Castelfranchi, and Falcone, trust has often been characterized as a decision to accept the risk associated with another [7], [9]. The current framework in figure 3 does not have a concept of trust, and only contains two actions. We now adopt the concept in figure 3 to a general action space and include two additional action choices- the action to "trust" and the action to "not trust." To "not trust" is to choose a subset of actions to pursue the independent reward. To "trust" is to choose a subset of actions to pursue the interdependent reward. The decision process becomes a two step decision process, where the first step is for the agent decide, at a particular state, is it more valuable to "trust" or "not trust" the other agent. This is a cognitive decision, where the state does not change. There is no physical movement of the agent. Then, depending on the trust/not trust choice, a subset of following actions are available. In choosing action to "trust", the agent has a subset of actions now available, and in choosing an action to "not trust," the agent has a different subset of actions available. Trust and not trust actions must influence availability of subsequent actions in order to differentiate trust and not trust action choices. Future work can focus on varying the probability of subsequent action choices. Figure 5-6 illustrates the new formulation.

Trust is formalized as an action. This is a distinction from previous trust-RL models [22]–[25]. These models represent trust as a parameter that is updated through interactions and other sources of information. We postulate that trust is a decision the agent faces, where it must estimate the value of trusting or not trusting in each state.

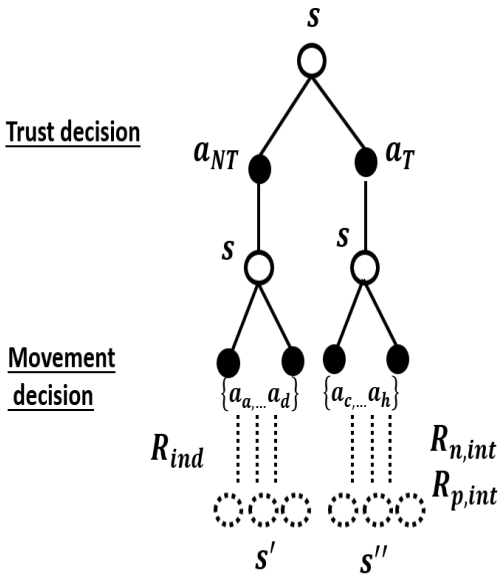

Figure 5: Formulation with trust as an action

Statespace: $s_i = \in S$
Actionspace: $a_i = \{a_{NT}, a_T, a_L, a_U\} \in A$
Transition function: T
Reward R, R_{ind}, $R_{n,int}$, $R_{p,int}$
π: ϵ - greedy; episodic
n: 2 agents; $\gamma \in$ [0,1] discount factor
β: Inherent risk-level
MDP = $\langle n, S, A, T, R, \gamma, \beta \rangle \rightarrow \{\pi_1^*, \pi_2^*\}$

Figure 6: MDP with trust as an action

3.4 Variance

The second problem with the standard RL formulation is how does the agent consider variance, due to the outcomes in the partner's behavior, in its decision-making process? Safe RL literature provides methods of addressing variance in return through changes in the objective function or exploration process. Other approaches may be to observe and estimate a policy of the other agent. Variance is included as part of the objective function [26], [28].

$$\pi^*(a|s) = \max_{\pi \in \Pi} (\mathbf{E}_\pi[G_t|s_t, a_t] + \begin{cases} \beta\sigma^2_{R,int} & ,if \quad a_t = T \\ 0 & ,if \quad a_t = NT \end{cases}) \tag{4}$$

Where, $\sigma^2_{R,int}$ is the variance of the historical interdependent outcomes. The variance is an estimation of the other agent's behavior. The agent maintains a record of the history of interdependent

outcomes. The variance of this distribution is the risk governed by the partner. If the firefighter scenario was a repeated game, Bob has attempted interacting with the firefighter x amount of times, sometimes successful, sometimes not. Bob approximates the risk governed by the firefighter through the variance of this distribution.

The parameter, β, serves to soften the component, and represents inherent risk preference of the agent. $\beta = 0$ denotes a risk-neutral agent, whose decision is driven only by maximum expected return. $\beta > 0$ denotes a risk-seeking agent, who has a propensity to accept the variance governed by the partner, which translates to placing additional value on trust actions. $\beta < 0$ denotes a risk-averse agent, who has a reluctance to accept the variance, which translates to removing value from trust actions.

The variance component only reinforces trust action choices. The partner is unreliable when the variance is large. Therefore a small β parameter can be selected to limit the influence the variance has on the value or utility for the action trust value. Most of the value will then come from the expected return from the interdependent goal. Conversely, the partner is reliable when the variance is small. Therefore a larger β parameter can be selected to increase the influence the variance has on the value or utility for the action trust value. Now, the value of the trust action will be bolstered by value of variance, in addition to the expected return.

3.5 Algorithmic Formulation

Two problems have been addressed: incorporating trust as an action and incorporating variance in the objective function. The agent now has to choose between depending on the partner and receiving an interdependent reward from a distribution, OR, acting alone and receiving a independently reward. With trust formulated as an action, the agent first determines if it is more valuable to not trust or to trust, which will dictate its subsequent movement actions. A variance component is added into the objective function so the decision to trust now considers the historical behavior of the partner.

We now have a more nuanced characterization of trust. The value to not trust is expected return from the independent goal. The value to trust is the expected return from the interdependent goal, with consideration to the risk governed by the partner. The value of trust is now meaningful in units of reward points. The value of trust is

now dependent on the state.

$$q_\pi(s, a_{NT}) = R_{ind} + \gamma \mathbf{E}_\pi[(G_{t+1}|s_{t+1}, a_{t+1})] \tag{5}$$

$$q_\pi(s, a_T) = \gamma \mathbf{E}_\pi[G_{t+1}|s_{t+1}, a_{t+1}] + \begin{cases} R_{p,int} + \beta\sigma^2_{R,int} & ,if \quad C = M \ \ and \ \ a_t = T \\ R_{n,int} + \beta\sigma^2_{R,int} & ,if \quad C = NM \ \ and \ \ a_t = T \\ R_{p,int} & ,if \quad C = M \ \ and \ \ a_t = NT \\ R_{n,int} & ,if \quad C = NM \ \ and \ \ a_t = NT \end{cases} \tag{6}$$

The agent now as a nuanced method to calculate the value of trust. At each state, it can determine the value to "not trust" versus to "trust". The value to "not trust" is the expected return in pursuing the independent reward, which only depends on the state, its own actions, and the independent reward. The value to "trust" is the expected return in pursuing the interdependent reward with consideration of the historical variance of the interdependency. The value of trust depends on the state, agent's own actions, the interdependent reward, and the behavior of the partner.

The on-policy method, SARSA(λ), is modified to include implement the formulation components specified above [21]. α is the learning rate of the agent, λ is the eligibility trace weight. The developed SARSA algorithm will be referred to as SARSA-Trust (SARSA-T) through the rest of this report.

$$q(s_t, a_t) \leftarrow q(s_t, a_t) + \alpha(R_{t+1} + \gamma q(s_{t+1}, a_{t+1}) - q(s_t, a_t) + (\beta\sigma^2_{R,int})) \tag{7}$$

Note that the above update is only used when a trust decision is made. When a "not trust" decision is made, the variance component is not included in the update.

3.6 Expanded state space

The next problem is to ensure the agent has accurate information from the environment to facilitate learning. The agent requires information on the benefits and consequences of its action choices, as related to the contingency. In this case, the contingency is that both agents end at the final

location together. Therefore, the agent's state space is defined as its own location and the location of the other agent.

4 Research Objectives and Questions

The theoretical formulation seeks to create a framework to study trust formation and collaboration between two agents. We use Interdependence Theory by Kelley and Thibault to characterize the a interaction between two agents, where trust and collaboration are possible. We incorporate the four trust model attributes of Castelfranchi and Falcone into a reinforcement learning framework. First, agents are motivated by their own reward functions. Second, the trust is formulated as a decision between an independent and interdependent reward. The final model characteristics, partner behavior prediction and subjectivity, are integrated into a modified objective function where a variance component. The value of trust at a particular state is the expected return from the interdependent reward plus the agent's subjective interpretation of the historical variance of the partner. The formulation within reinforcement learning allows us to investigate trust as a learned behavior through a sequential decision-making process. Trust is learned through repeated interactions with the partner. Feedback from decisions to trust or not trust are delayed.

Aspects of the theoretical formulation enable us to pursue two research objectives. 1) To determine under what conditions is collaboration between two agents achieved. 2) To determine what the value of trust reveals about the interactions under these conditions. Through analysis of the framework's rewards, agent subjectivity, and behavior, we aim to understand trust formation and collaboration between two agents.

4.1 Research Questions

#1: Impact of SARSA-Trust. How does the new theoretical formulation affect trust and collaboration between the two agents? The research question investigates the impact of each new component of the theoretical formulation, to trust and collaboration between the two agents.

#2: Rewards. How do different values in the rewards affect trust and collaboration between the two agents? The research question investigates the role of the reward components- the independent reward, R_{ind}, the positive interdependent reward, $R_{p,int}$, and the negative interdependent

reward, $R_{n,int}$, in trust formation and collaboration.

#3: Risk preference. How does inherent risk preference effect trust and collaboration between the two agents? The research question assesses the role of subjective considerations of risk preference in trust formation and collaboration.

#4: Partner behavior. How does partner's behavior effect trust and collaboration between the two agents? The research question analyzes the impact of varying partners in trust formation and collaboration.

5 Experimental Design

We will utilize a simulated environment to investigate the research questions. The gridworld environment is a useful tool to explore the representation of an MDP [21]. A gridworld's interpretability and general applicability make it an excellent place to begin the exploration of trust between agents within an RL framework. Particularly, gridworld allows for the testing of sequential decision-making, where immediate trust/not trust decisions can have long term benefits and costs. Figure 7 illustrates the experimental design. A four-by-four Cartesian gridworld is utilized. Two agents are included in the experiment. The objective for each agent is to maximize their individual reward, which is achieved through finding the shortest path to one of two terminal locations, location 0 or location 15. The state space of the agent is a tuple comprising of the location of agent 1 (grids 0 through 15) and the location of agent 2 (grids 0 through 15). There are six possible actions available for each agent: to "not trust", to "trust" move left, move up, move right, or move down.

Each agent is given its own reward function that has a independent reward, R_{ind}, a positive interdependent reward, $R_{p,int}$, and a negative interdependent reward, $R_{n,int}$. If the agent arrives at location 0, it will receive the independent reward, R_{ind}; the reward is independent of the other agent's actions. The interdependent reward is conditioned on the behavior of both agents. If both agents arrive at location 15, both agents will receive the positive interdependent reward, $R_{p,int}$. If one agent goes to location 15 and the other agent goes at location 0, the former will receive the negative interdependent reward, $R_{n,int}$, as a failed attempt to collaborate. Finally, both agents receive a "-1" point for every step they take, to emphasize the shortest path to the terminal states. Figure 8 summarizes the reward structure in game theoretic matrix form. Summarizing the reward

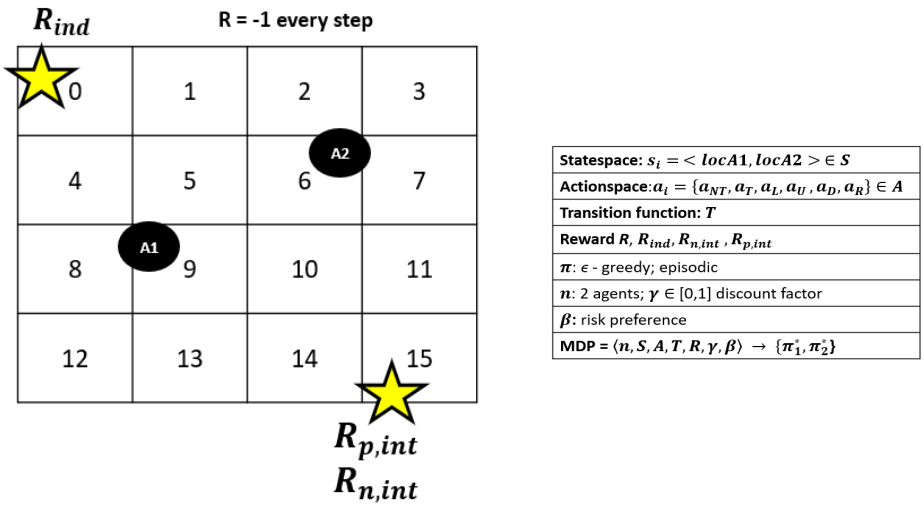

Figure 7: Experimental formulation

structure in a matrix highlights the interdependency between the two agents. The matrix also illustrates this how this can be seen as a simple coordination game between the two agents.

		Agent 2	
		S0	S15
Agent 1	S0	R_{ind} / R_{ind}	$R_{n,int}$ / R_{ind}
	S15	R_{ind} / $R_{n,int}$	$R_{p,int}$ / $R_{p,int}$

Figure 8: Rewards in matrix form

Figure 9 illustrates the two decisions that the agent makes during each iteration. The first decision is the trust decision. Following an ϵ-greedy policy, the agent decides either to "trust" or "not trust" the other agent. There is no change in the state; no physical movement of the agents. Furthermore, the trust decision will impact the availability of actions during the second decision. If the agent decides to "trust", then only a subset of actions pursuant of the interdependent reward is made available- move right and move down. If the agent chooses to "not trust", then only a subset of actions pursuant of the independent reward is made available- move left and move up.

A algorithmic cycle is a cycle of two decisions that the agent goes through; first a trust decision,

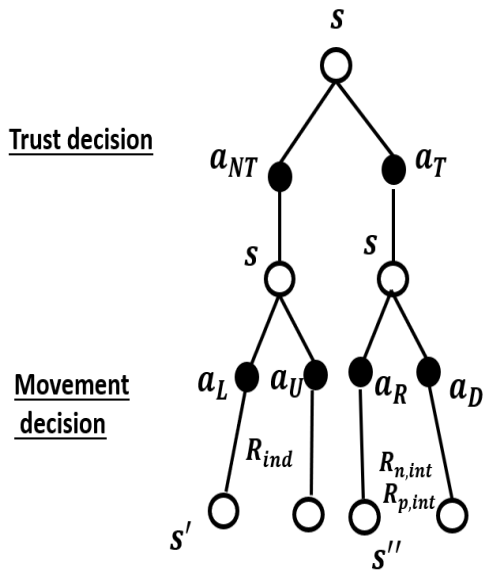

Figure 9: Agent decision tree in Gridworld

then a movement decision, after which results in one physical movement step. An agent will make any number of algorithmic cycles to move to a terminal location, as defined as location 0 or location 15. Once both agents reach a terminal location, the episode concludes. The action-values calculated during that episode is transferred to the next episode. Both agents start at new initial locations and begin a new episode to alternate movement to a terminal location. There is a specified number of episodes to allow both agents a specified number of repeated interactions and learn how to optimally behave in given conditions.

5.1 Outcome Variables

Two sets of outcome variables are used: final outcomes and trust-values. Final outcome metrics are borrowed from game theoretic approaches [32] and give insight into team and collaborative performance. What goals do the agents learn to pursue? An episode concludes with either agent finishing at a particular terminal location. Figure 10 illustrates the four possible outcomes. The fraction of outcomes, out of total tested episodes, are indicative of what policies the agents have learned under given conditions. If both agents terminate at location 0 (S0S0), this indicates both agents have chosen to act independently for that episode. If both agents choose terminate at location 15 (S15S15), both agents have chosen to collaborate. This is indicative of mutual collaboration

and a successful interaction as both agents will have obtained the positive interdependent reward. If agent 1 terminates at location 0 and agent 2 terminates at location 15 (outcome S0S15), this is indicative of a sub-optimal decision or lack of coordination. This is also true when agent 1 terminates at location 15 and agent 2 terminates at location 0 (outcome S15S0).

		Agent 2	
		S0	S15
Agent 1	S0	Mut. Defection (S0-S0)	Miscoord. (S0-S15)
	S15	Miscoord. (S15-S0)	Mut. Collab. (S15-S15)

Figure 10: Final Outcomes

Trust-values give insight into the trust dynamics between the agents and why the agents behave the way they do. Equations 5-7 enable the calculation of trust values by state. Specifically, we look at the expected value of trust minus the expected value of not trust for each state. If the difference is positive, this indicates the state is beneficial for "trust". If the difference is negative, this indicates the state is beneficial to "not trust".

$$\Delta = q_\pi(s_t, a_T) - q_\pi(s_t, a_{NT}) \tag{8}$$

When comparing the effect of different treatments on trust values, the percent of states that favor trust will be used as a summary metric. In any particular state, the agent is going to find it more valuable to "trust" or to "not trust". The share of states that the agent finds it more valuable to "trust" is a summary of the impact of the treatment on trust.

$$S_{\%,T} = \frac{\#\, of\ states\ where\ \Delta > 0}{total\ \#\, of\ states} \tag{9}$$

5.2 Demonstrated performance

The following subsection demonstrates the developed framework and how the outcome variables can be utilized to analyze performance.

In the following situation, two independent learning agents are given the same reward function and developed SARSA-Trust algorithm. The reward components are set at, $R_{ind} = 9.5$, $R_{p,int} = 20$, and, $R_{n,int} = -1$. Both agents are risk-neutral agents at $\beta = 0$. Therefore, they do not consider variance in their objective function. Both agents are unsuccessful in achieving mutual cooperation. Figure 11 illustrates their final outcomes. Out of 1,000 episodes, 69.68% of the episodes results in both agents terminating at location 0. Therefore, 69.68% of the time, both agents acted independently or mutual defection. 15.68% of the episodes resulted in both agents terminating at location 15. Therefore, 15.68% of the time, both agents collaborated successfully or mutual collaboration. 7.30% of the episodes resulted in agent 1 going to location 0 while agent 2 terminated at location 15, and 7.34% of the episodes resulted in the opposite outcome. These illustrate mis-coordinated events among the two agents.

		Agent 2	
		S0	S15
Agent 1	S0	69.68 (1.65)	7.30 (1.14)
	S15	7.34 (1.40)	15.68 (2.34)

* both agents are risk neutral (β =0)
* out of 1,000 episodes
* standard deviation in paranthesis

Figure 11: Final Outcomes for demonstration, Risk-Neutral Agents

Figure 12 illustrates the plot of the trust-values that Agent 1 has learned. Agent 1 calculates its value to "not trust" based on its proximity to the independent reward, location 0. Agent 1 calculates its value to "trust" based on its proximity to the interdependent reward at location 15, Agent 2's proximity to location 15, and Agent 2's prior behavior. The difference between the q-values dictate which outweighs the other. A positive difference notes that it is more valuable to

"trust" than to "not trust" at a particular state.

The y-axis is the delta trust metric, in units of expected return. The x-axis annotates the number of steps Agent 1 is away from location 15. For example, if Agent 1 is one step away from state 15, Agent 1 is located at position 11 or position 14. Since this graph is from the perspective of Agent 1, steps one and six are not shown for the agent. These indicate Agent 1 has reached a terminal location and the episode is complete. The functions are categorized by color based on how many steps Agent 2 is from location 15. For example, if Agent 2 is zero steps away from state 15, Agent 2 is located at state 15.

Overall, as the number of steps away from location 15 increase, Agent 1 gets further away from the interdependent reward and closer to the independent reward. Following this trend, the value of trust decreases.

The value of trust is the highest when Agent 2 is located at terminal location 15. The only time Agent 1 finds value to trust Agent 2, is when Agent 2 is actually at location 15 guaranteeing a successful interaction outcome. This makes sense as Agent 1 is a risk neutral agent and does not accept any risk. Therefore, only in situations when the interdependent reward is guaranteed does the agent find value in the action to "trust". Once Agent 1 is four and five steps away from location 15, while Agent 1 is at location 0, it is no longer valuable for agent 1 to "trust" agent 2.

To summarize the trust impact on this scenario, approximately 2.72 - 3.26% of the state space favors trust. From approximately three percent of the states that Agent 1 finds itself in, where it will choose to "trust" over to "not trust" thereby explaining the high rates of acting independently.

Figure 13 illustrates the outcomes when both agents are risk-seeking agents at $\beta = 0.1$. Now both agents consider and magnify variance as part of their objective function. Only 6.24 % of the time, both agents act independently and pursue the reward at location 0. 62.54 % of the time, both agents learn to mutually cooperate, and successfully obtain the positive interdependent reward at location 15. 14.51-16.71% of the time results in mis-coordinated outcomes on the part of both agents. The final outcomes demonstrate an increase in mutual collaboration, when both agents are risk-seeking.

Figure 14 illustrates the action value plot of Agent 1, when both agents are risk-seeking at $\beta = 0.1$. It is helpful to compare the changes from Figure 12 (risk-neutral agents) and Figure 14 (risk-seeking agents). Overall, the is more variability introduced in the values, due to the variation

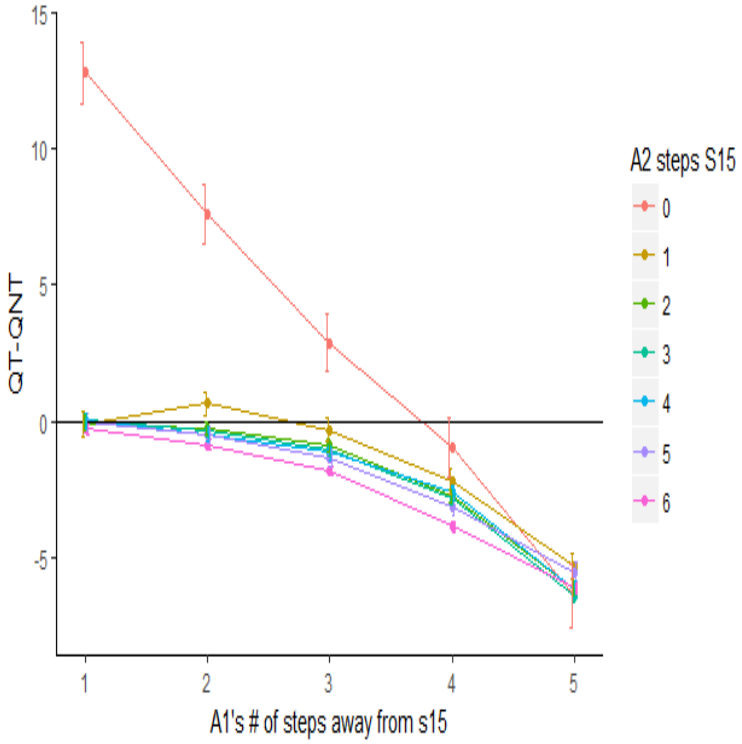

Figure 12: Agent 1's trust in agent 2, Risk-Neutral Agents

in the objective function. When Agent 2 is located at terminal location 15, the value of of the action, to "trust", does not change between risk-neutral and risk-seeking agents. Once Agent 2 is located at the terminal location 15 the positive interdependent reward is guaranteed, the variance and risk is minimized, and therefore there is no risk to assume. A risk-seeking Agent 1 starts to assume risk at other states. There is now a visible change when Agent 2 is one step away. A risk-neutral Agent 1 will not find it valuable to "trust" whenever Agent 2 is one step away. However, a risk-seeking Agent 1 does find it valuable to "trust" when Agent 2 is one step away at certain states. To summarize Figure 14, approximately 37.38-38.92% of the state space favors trust. From approximately 38% of the states that Agent 1 finds itself in, it will choose to "trust" over to "not trust", which leads to sustainable mutual collaboration results.

The increase state space coverage supporting the trust action results in increase mutual collaboration rates. However, a comparison of the final outcomes of in Figure 11 (risk-neutral agents) and Figure 13 (risk-seeking agents) shows that risk-seeking agents result in an increase in miscoordinated outcomes. On one hand, the risk that the agents assume increases in mutual collaboration. However, increases in risk assumption also result in potential failures in interaction. These

		Agent 2	
		S0	S15
Agent 1	S0	6.24 (4.61)	14.51 (6.71)
	S15	16.71 (8.18)	62.54 (4.12)

* both agents are risk seeking (β =0.1)

* out of 1,000 episodes

* standard deviation in paranthesis

Figure 13: Final Outcomes for demonstration, Risk-Seeking Agents

failures can be seen by the increases in mis-coordinated outcomes.

5.3 Tests

For the first research question, the impact of theoretical formulation, we test the addition of each component of the framework. The purpose of the tests are to determine the impact of the addition of each developed component to the overall collaboration between the two agents. The tests and hypotheses are outline in 15.

For test #1, the agents will be given the standard SARSA(λ) algorithm, to determine baseline performance. Can the agents learn to collaborate with a standard RL algorithm? For test #2, agents are given additional information about their partner's current location, in the form of an expanded state space. Does giving the agent the current location of their partner help the agents learn to collaborate? For test #3, the agents' state space returns back to only their self-location, but the action space now includes the actions to "not trust" and to "trust". Does the trust as an action formulation, or awareness of their trust actions, give the agents the capability to collaborate? For test #4, the agents are given the expanded state space and trust actions. For test #5, the agents are given the full SARSA-Trust algorithm which includes the expanded state space, the trust actions, and the inclusion of variance in the objective function. Does it take the modification of the objective function and risk-seeking behavior for the agents to learn to collaborate?

Both agents will be given the same reward function of $R_{ind} = 9.5$, $R_{p,int} = 20$, and, $R_{n,int} = -1$ for all the tests. A simple 2x2 matrix game can give overall expected values of either decision,

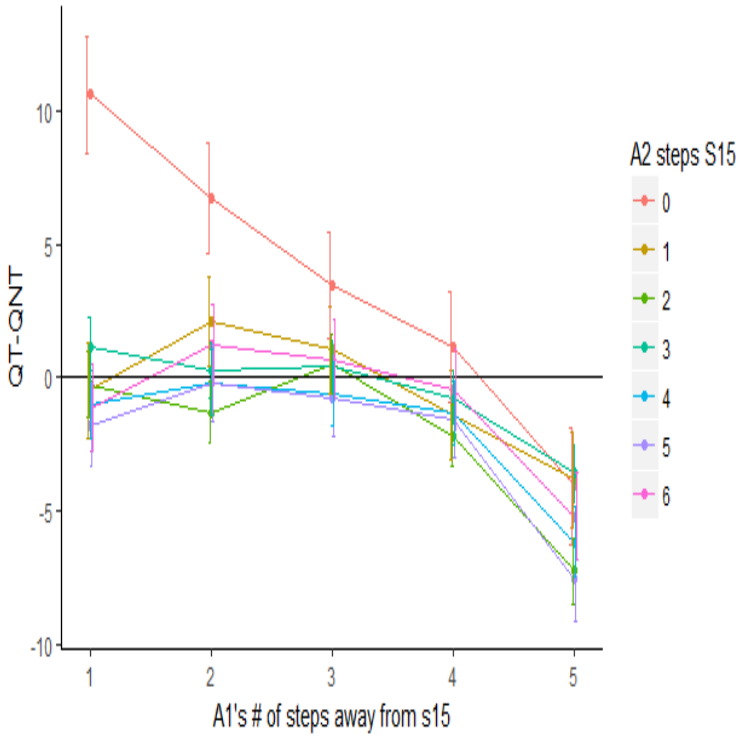

Figure 14: Agent 1's trust in agent 2, Risk Seeking Agents

assuming the other agent behaves in a probability of 50/50. If Agent 2 is assumed to act in a 50/50 manner, Agent 1's expected value in acting independently and collaborating are both 9.5.

$EV(act\,ind) = R_{ind} = 9.5$

$EV(to\,collab) = P_{A2,S0}R_{n,int} + P_{A2,S15}R_{p,int} = (0.5)(-1) + (0.5)(20) = 9.5$

Research question #2, Rewards, will explore if there are situations where collaboration can occur among risk-neutral agents through value changes in the reward structure. To test research question # 2, several reward values will be evaluated. The tests, experimental settings, and hypotheses are articulated in 16. In test #1, the independent reward,R_{ind}, is higher than the positive interdependent reward$R_{p,int}$. In the subsequent tests, the independent reward,R_{ind}, will be lowered with a constant positive interdependent reward$R_{p,int}$. This is to determine when the agent makes trades between the independent and interdependent reward, and based on expected value and variance. The final set of tests are to determine how the agents decision-making adopts to changes in the negative interdependent reward, $R_{n,int}$.

Research question #3 explores the inherent risk preference of the agents. Figure 17 summarizes

Case	Setting	Hypothesis
SARSA(λ)	$\langle n, S, A, T, R \rangle$	MD > 50 %, $S_{\%,T}$ < 20 %
Partner awareness (S)	$s_i < locA1, locA2 > \in S$	MD > 50 %, $S_{\%,T}$ < 20 %
Trust as action (A)	$a_i = \{a_{NT}, a_T, a_{L,} a_U, a_R, a_D\} \in A$	MD > 50 %, $S_{\%,T}$ < 20 %
Partner awareness (S) + trust as action (A)	$\langle n, S, A, T, R \rangle$	MC = 50 %, $S_{\%,T}$ = 50 %
Partner awareness (S) + trust as action (A) + Risk-seeking (β)	$\langle n, S, A, T, R, \beta \rangle$	MC > 50 %, $S_{\%,T}$ > 50 %

MD (mutual defection), MC (mutual collaboration), $S_{\%,T}$ (percent of state space favoring trust).
1,000 episodes in experiment, with $R_{ind} = 9.5, R_{n,int} = -1, R_{p,int} = 20$.
Statistical validity through pairwise t-test at 95% confidence interval.

Figure 15: Tests and hypotheses, RQ #1 (Impact of formulation)

the tests for this research question. The risk preference parameter, β, for both agents will be incrementally increased so the agents will be made incrementally more risk-seeking. Then, the parameter will be incrementally decreased (more negative), so the agents will be made incrementally more risk-averse.

Research question #4, partner behavior,investigates four different behaviors that Agent 1 must interact and adapt to. The tests, experimental settings, and hypotheses are summarized in figure 18. In test #1, Agent 1 interacts with a "cooperative" Agent 2. A "cooperative" Agent 2 is given a reward function that only leads to location 15. This results in greater than 90% of Agent 2 terminations at location 15. In test #2, Agent 1 interacts with a "non-cooperative" Agent 2. A "non-cooperative" Agent 2 is given a reward function that only leads to location 0. This results in greater than 90% of Agent 2 terminations at location 0. In test #3 and test #4, Agent 1 interacts with an "unbiased" Agent 2. An "unbiased" Agent 2 is given the same reward function as Agent 1 and both agents must determine if collaboration is beneficial. We test a "unbiased, risk-neutral" Agent 2, and a "unbiased, risk-seeking Agent 2".

The outcome variables for each test within each research question are as previously specified-final outcomes and the delta trust values. The outcome variables will be used to determine the impact of the treatment on collaboration and separately, trust, between the two agents. Each test will be executed over 1,000 episodes. Ten iterations of each test will be conducted to determine standard errors. Pairwise t-test will be utilized at a 95 % confidence interval to evaluate each

Case	Setting	Hypothesis
$R_{ind} > R_{int}$	$R_{ind} = 22, R_{n,int} = -1, R_{p,int} = 20$	MD > 50 %, $S_{\%,T} < 20$ %
$R_{ind} < R_{int}$	$R_{ind} = 9.5, R_{n,int} = -1, R_{p,int} = 20$	MD > 50 %, $S_{\%,T} < 20$ %
	$R_{ind} = +1, R_{n,int} = -1, R_{p,int} = 20$	MC > 50 %, $S_{\%,T} > 50$ %
	$R_{ind} = -1, R_{n,int} = -1, R_{p,int} = 20$	MC > 50 %, $S_{\%,T} > 50$ %
$R_{ind} < R_{int}$	$R_{ind} = 9.5, R_{n,int} = +1, R_{p,int} = 20$	MC > 50 %, $S_{\%,T} > 50$ %
$R_{ind} > R_{int}$	$R_{ind} = 9.5, R_{n,int} = -10, R_{p,int} = 20$	MD > 50 %, $S_{\%,T} < 20$ %

MD (mutual defection), MC (mutual collaboration), $S_{\%,T}$ (percent of state space favoring trust).
1,000 episodes in experiment, with specified rewards.
Statistical validity through pairwise t-test at 95% confidence interval.

Figure 16: Tests and hypotheses, RQ #2 (Rewards)

Case	Setting	Hypothesis
Risk-averse	$\beta < 0$	MD > 50 %, $S_{\%,T} < 10$ %
Risk-neutral	$\beta = 0$	MD > 50 %, $S_{\%,T} < 20$ %
Risk-seeking	$\beta > 0$	MC > 50 %, $S_{\%,T} > 50$ %

MD (mutual defection), MC (mutual collaboration), $S_{\%,T}$ (percent of state space favoring trust).
1,000 episodes in experiment, with $R_{ind} = 9.5, R_{n,int} = -1, R_{p,int} = 20$.
Statistical validity through pairwise t-test at 95% confidence interval.

Figure 17: Tests and hypotheses, RQ #3 (Risk Preference)

hypothesis.

6 Results

6.1 RQ #1: Impact of Formulation

The first research question asks how does the SARSA-Trust algorithm effect trust formation and collaboration between the two agents.

Figure 19 summarizes the effect of adding different components of the algorithm on collaboration between the two agents. The agents are posed with a situation to determine if it is advantageous to pursue a independent reward at location 0 of 9.5 points, or to work together for an interdependent reward at location 15 of 20 points. When both agents are given the standard SARSA(λ) algorithm,

Case	Setting	Hypothesis
A2 "non-cooperative"	A2[$R_{S0} = 10, R_{S15} = 0$], $\beta_2 = 0.0$	MD > 50 %, $S_{\%,T} < 20$ %
A2 "cooperative"	A2[$R_{S0} = 0, R_{S15} = 10$], $\beta_2 = 0.0$	MC > 50 %, $S_{\%,T} > 50$ %
A2 "Unbiased, risk-neutral"	A2[$R_{ind} = 9.5, R_{n,int} = -1, R_{p,int} = 20$], $\beta_2 = 0.0$	MC = 50 %, $S_{\%,T} = 50$ %
A2 "Unbiased, risk-seek."	A2[$R_{ind} = 9.5, R_{n,int} = -1, R_{p,int} = 20$], $\beta_2 = 0.01$	MC > 50 %, $S_{\%,T} > 50$ %

MD (mutual defection), MC (mutual collaboration), $S_{\%,T}$ (percent of state space favoring trust)
1,000 episodes in experiment, based on preliminary results.
Statistical validity through pairwise t-test at 95% confidence interval.

Figure 18: Tests and hypotheses, RQ #4 (Partner Behavior)

	Treatment effects			
	s0s0	**s0s15**	**s15s0**	**s15s15**
Tests				
SARSA(λ)	84.85	6.60	7.80	0.68
(sd)	(1.23)	(1.05)	(0.85)	(1.75)
Partner awareness (S)	83.48	7.40	8.03	1.03
(sd)	(1.74)	(1.48)	(1.20)	(2.47)
Trust as action (A)	60.63	17.03	17.55	4.70
(sd)	(1.74)	(1.48)	(1.20)	(2.47)
Part. awar. (S) + trust (A)	69.68	7.34	7.30	15.68
(sd)	(1.65)	(1.40)	(1.14)	(2.34)
Part. awar. (S) + trust (A) + risk seek. (β) (sd)	6.24	14.51	16.71	62.54
	(4.61)	(6.71)	(8.18)	(4.12)

Figure 19: Impact of addition of components to collaboration

both agents adopt a policy of mutual defection 84.85 % of the time. Between 6.6-7.8 % of the time, the agents miscoordinate and randomly terminate at location 15. Mutual cooperation is never achieved through a standard SARSA(λ) algorithm.

With test #2, the agents are given awareness of their partner's current location, by including that information in an expanded state space. There are no significant changes to final outcomes. With test #3, the agents do not have the partner awareness in their expanded state space, but are given the trust actions. This leads to mutual defection rates decreasing to 60.63 %. However, mutual cooperation rates only increase to 4.70 %, while miss-coordinated efforts increased to approximately 17.0 %. The stark increase in mis-coordination rates illustrates the need for additional information on the partner to reduce the miscoordinated events.

Test #4, results reveal that expanding the state space and including trust as an action work effectively together. Mutual defection rates are reduced by 69.68 % and mutual cooperation rates increase to 15.68 %. The mis-coordination rates marginally change. The expanded state space gives the agents the right information to effectively obtain the interdependent reward. The trust action formulation ensures the reinforcement (or feedback) remains distinct; independent rewards reinforce only the action to not trust and associated movement actions, while interdependent rewards reinforce only the action to trust and associated movement actions. However, agents will act indepdently the majority of the time.

Test #5 is the the full SARSA-Trust algorithm with risk seeking agents at $\beta = 0.1$, which includes the variance component in the objective function. SARSA-Trust leads to a mutual defection decrease to 6.24 % and a mutual cooperation increase to 62.54 %. It is not until risk-seeking behavior does collaboration occur.

In summary, when two agents are posed with a choice between a independent and interdependent reward, the standard SARSA algorithm demonstrates that the agents will learn to act independently. The first research question seeks to determine if the proposed formulation can improve mutual collaboration. It is shown that in order for the agents to collaborate, at a minimum the agents must have a method for comparing the independent option with the interdependent option (trust as an action). Integrating more information into the state space to help obtain the interdependent reward leads to a reduction in miscoordination events and improving mutual collaboration. However, the inclusion of the variance component in the objective function and risk-seeking behavior provides the agents is the most significant capability that leads to mutual collaboration.

6.2 RQ #2: Rewards

The results of research question # 1 suggest the importance of risk-seeking behavior for mutual collaboration. This motivates research question # 2, how do rewards effect trust formation and collaboration between the two agents? Can we find ways for risk-neutral agents to collaborate through different values in the reward components. To note, the following tests are conducted with risk-neutral agents only.

Figure 20 summarizes the final outcomes of the research question. When the independent reward is 22.0, while the positive interdependent reward is 20.0 points, mutual defection is achieved

	Outcomes			
	s0s0	s0s15	s15s0	s15s15
Tests				
Rind = 22.0	76.41	6.86	6.56	10.17
(sd)	(1.82)	(1.51)	(2.46)	(3.00)
Rind=9.5	70.29	7.75	7.55	14.41
(sd)	(1.63)	(1.58)	(3.07)	(2.79)
Rind = +1.0	46.84	12.93	12.12	28.11
(sd)	(0.93)	(1.97)	(0.92)	(2.60)
Rind = -1.0	25.07	11.84	11.42	51.67
(sd)	(0.95)	(0.96)	(1.06)	(1.57)
Rnint=+1.0	63.56	9.86	11.89	14.69
(sd)	(2.56)	(1.85)	(2.30)	(2.45)
Rnint=-10.0	71.72	7.46	6.32	14.50
(sd)	(1.25)	(1.78)	(2.25)	(3.15)

Figure 20: Final Outcomes for changing independent reward

76.41 % of the time. If the independent reward is lowered to 9.5, mutual defection rates slightly decrease. If the independent reward is 1.0 points, mutual defection rates of 46.84 % can be achieved. Additionally, if the independent reward is made -1.0 point, mutual collaboration rates increase to 51.67 %. These series of tests demonstrate that the agents will learn a policy to the higher expected return. In most cases, the independent reward is the higher expected return.

The final two tests explore the impact when the negative interdependent reward is changed. For these two tests, the independent reward is set at 9.5 points. In comparison to test #2 ($R_{ind} = 9.5$, raising the negative interdependent reward to +1.0 points does result in a decrease in mutual defection rates from 70.29 % to 63.56 %. This make sense as the expected value of the interdependent reward is increased slightly. To note, most of the complementary increase in outcomes is in miscoordinated events rather than mutual cooperation. The final test explores raising the negative interdependent reward to -10.0 points. The changes are not statistically significant than the negative reward set at -1.0 points. This may indicate the model is not effective in articulating magnitude differences in costs.

Figure 21 summarizes the results of the delta trust metric under each treatment. For independent reward values of 22.0 and 9.5 points, the share of states that favor trust range from approximately 15.0-19.0 %. There are few states where it is more valuable to "trust" than to "not trust". This illustrates why the agents learn policies to act independently the majority of the time. It is not until the independent reward is set at -1.0, that approximately half of the state space favors trust. The agent finds it valuable to "trust" from half of the states. This test results in

95% CI, share of stateSpace favoring trust

Tests	LB	UB
Rind = 22.0	15.84	18.77
Rind=9.5	16.52	19.50
Rind = +1.0	31.49	35.15
Rind = -1.0	47.32	51.20
Rnint=+1.0	19.21	22.35
Rnint=-10.0	16.59	19.58

Figure 21: Trust coverage

policies towards the interdependent reward, and resulting in mutual collaboration rates.

Overall, the agents will learn the optimal policies to pursue the goals of higher expected value. Mutual collaboration is achieved between the agents when the higher expected value shifts towards the interdependent reward. Correspondingly, the share of states that value trust also increase. The tests demonstrate collaboration is not synonymous with trust. High values of trust can lead to mutual cooperation as well as increased mis-coordinated rates. Moreover, what informs the agent's decision to "trust" is only expected return from the interdependent reward. The agent assumes the risk based only on limited information (expected return). What if the agent is given more information about the interdependent reward? In addition to expected return of the interdependent reward, what if the agent is given the variance of the interdependent reward? Does this additional information yield better decision-making?

Mutual collaboration only occurs when the independent reward is negative. These results suggest that mutual collaboration among risk-neutral agents is attained in the absence of an independent reward. This further reinforces the notion that some level of risk-seeking behavior is required for mutual collaboration.

6.3 RQ #3: Risk Preference

The results of research question #2 suggest some level of risk-seeking behavior is required for mutual collaboration. This motivates the third research question to investigate the impact of inherent risk preference of the agent on trust formation and collaboration between the two agents.

Figure 22 summarizes the final outcomes for the research question. The agents become in-

	Outcomes			
beta	**s0s0**	**s0s15**	**s15s0**	**s15s15**
-1.0	82.78 (0.92)	8.24 (0.56)	8.11 (0.88)	0.87 (0.21)
-0.1	83.51 (0.93)	6.63 (0.56)	6.84 (0.57)	3.02 (0.60)
-0.01	79.47 (1.11)	6.14 (1.17)	6.17 (2.38)	8.22 (2.65)
-0.001	72.22 (2.67)	7.65 (1.86)	6.81 (1.43)	13.32 (3.34)
-0.005	76.92 (1.77)	6.97 (1.68)	6.97 (1.91)	9.14 (2.71)
0	70.02 (1.79)	7.68 (2.32)	8.08 (2.43)	14.22 (2.67)
0.001	67.42 (1.31)	7.79 (2.81)	8.19 (2.32)	16.60 (3.63)
0.005	48.04 (5.78)	11.55 (5.72)	12.59 (6.14)	27.82 (4.29)
0.01	23.06 (5.84)	13.44 (9.03)	19.61 (7.65)	43.89 (4.49)
0.1	6.24 (4.61)	14.51 (6.71)	16.71 (8.18)	62.54 (4.12)
1.0	2.70 (1.50)	13.76 (3.27)	14.70 (2.90)	68.84 (4.44)

* β parameter for both agents

* reported in percent

* ten iterations of 1,000 episodes each

Figure 22: Final outcomes for varying inherent risk preference

creasingly risk-seeking agents, through a β parameter increase from 0 to 1.0. By β=0.1, a mutual collaboration rate of 62.54 % is achieved. Mutual defection rates are significantly reduced, and as low as 6.24 %. As the β parameters increase, more variance is included to the expected return. If $\beta \geq 1.0$, the variance will dominate the calculated values.

A trade-off is identified. As mutual cooperation increases, mis-coordinated events also increase. These are sub optimal decisions. By β=0.1, 14.51 % of the time Agent 1 went to location 15 despite Agent 2 going to location 0. Agent 1 can be seen as having over-trust in Agent 2 in these instances. 15.51 % of the time, Agent 1 trusts Agent 2, to its own detriment.

As the β parameters become increasingly negative, the agents becoming increasingly risk-averse. The agents show an increased preference towards the independent reward. The mutual defection rates increase. Levels of mutual collaboration are effectively reduced. However, overall the approach does not effectively demonstrate risk aversion. The miscoordination rates can be interpreted as

under-trust or missed opportunities. At at risk level of β=-0.1, 6.63 % of instances occur where Agent 1 missed an opportunity to collaborate with Agent 2.

β	95 % CI	
-1.0	0.14%	0.30%
-0.1	0.40%	0.65%
-0.01	1.95%	2.46%
-0.005	3.96%	4.66%
-0.001	6.91%	7.81%
0.0	2.72%	3.26%
0.001	7.81%	8.68%
0.005	15.24%	16.40%
0.01	25.97%	27.37%
0.1	37.38%	38.92%
1.0	49.24%	50.82%

* ten iterations of 1,000 episodes eacb

Figure 23: State space coverage

Figure 23 illustrates the impact of varying inherent risk preference on trust. As both agents become increasing risk-seeking, more of their state space favors trust actions. On the other hand, as both agents become increasingly risk-averse, less of their state space favors trust actions.

The results also highlights the gap between trust and collaboration. Trust does not guarantee collaboration [7]. This is due to the role of subjectivity in trust calculations. Agents with risk-seeking levels of $\beta = 0.1$ have approximately 38.0 % of their state space favoring trust. This does not translate to an approximate 76% of mutual collaboration. The two agents under these risk-seeking levels achieve a mutual collaboration rate of 62.54 %, while 14.51-16.71 % are miscoordinated or over-trust outcomes. This gap is just one way collaboration and trust are separate constructs.

6.4 RQ #4: Variation in partner behavior

The fourth research question investigates how the partner's behavior effect trust formation and collaboration between the two agents. The previous research question (RQ # 3), identifies the trade-off between risk-seeking behavior, trust, and collaboration. How can agents mitigate the detrimental effects of over-trust outcomes?

Figure 24 illustrates the first scenario where Agent 1 encounters a "cooperative" Agent 2, an Agent 2 that follows a policy only to location 15. This results in a Agent 2 that travels to location 15 greater than 80% of the time. When Agent 1 is risk-neutral, with a $\beta = 0$, mutual collaboration is not achieved. Agent 1's rates of acting independently are indicated by the rates of S0S0 plus

	Outcomes- A2 "cooperative"			
β_1	s0s0	s0s15	s15s0	s15s15
0.0	22.49	55.12	3.37	19.02
	(1.47)	(2.80)	(0.73)	(2.87)
0.005	18.10	45.36	6.71	29.83
	(2.64)	(4.92)	(2.82)	(3.67)
0.001	21.99	53.62	4.06	20.33
	(1.17)	(2.86)	(1.22)	(2.33)
0.01	10.69	26.24	13.30	49.77
	(3.61)	(8.06)	(4.58)	(7.01)
0.1	4.43	19.27	30.13	39.63
	(2.51)	(6.43)	(8.41)	(11.93)

* reported in percent

* ten iterations of 1,000 episodes eacb

Figure 24: Final Outcomes, Interaction with "Cooperative" A2

S0S15. As Agent 1 increases in inherent risk-seeking behavior, the rate of mutual collaboration increases. By a risk-seeking level of $\beta_1 = 0.01$, mutual collaboration rates of 49.77 % are achieved. However, there is a trade-off in S15S0 events as well. As mutual cooperation increases, the number of instances where Agent 1 goes to location 15 without Agent 2 also increases. This is indicative of over-trust in Agent 2. At β_1=0.01, Agent 1 goes to location 15 at a rate of 13.30 % of the time to its detriment, indicating over-trust behavior. If that risk-preference is increased to β_1=0.1, Agent 1 goes to location 15 at a rate of 30.13 % of the time.

Choosing the optimal risk-seeking parameter is dependent on the partner, situation, and rewards that the agent faces. It can be seen as an optimization problem with two objectives. 1) What is the minimum amount of mutual collaboration that I want to achieve?, 2) What is the maximum amount of negative outcomes from over-trust that I am willing to accept? We must find the optimal risk-seeking parameter beta that satisfies these objectives in the current situation. Future work will focus on more effective optimization strategies. Currently, the parameter is selected visually. In this situation, when Agent 1 encounters a "cooperative" Agent 2, a suitable risk-seeking level is β_1=0.01.

Figure 25 illustrates when Agent 1 is risk neutral, it learns a policy towards the independent reward (S0S15) outcomes. Once Agent 1 is risk-seeking to the appropriate level of β_1=0.01, it is able to learn a policy to mutual collaborate with a "cooperative" Agent 2.

When examining the action-value plots for a risk-neutral Agent 1, 2.04-3.43% of the state space

favors trust. The risk-seeking Agent 1 has trust favorable coverage expanded to 36.07-37.13% of the state space, enabling it to learn a mutually collaborative policy.

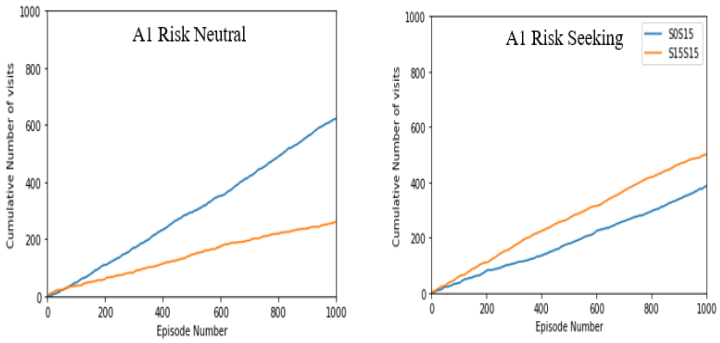

Figure 25: Learning performance of A1 interacting with "Cooperative" A2

Figure 26 illustrates the second scenario where Agent 1 interacts with a "non-cooperative" Agent 2, an Agent 2 that follows a policy only to location 0. This results in a Agent 2 that travels to location 0 greater than 80% of the time. In this scenario, Agent 1 must demonstrate it can learn a optimal policy to act independently. A new inherent risk level is required for Agent 1 to adapt optimally to the new type of Agent 2 behavior. An Agent 1 that is optimized to the previous scenario, β_1=0.01, will still learn a policy to act independently. However the performance is mediocre, as over-trust rates are very high (46.40 %). To improve these metrics, Agent 1's risk level needs to be adjusted.

A risk-neutral Agent 1 provides considerable improvement, where mutual defection rates are now 71.04 %. Agent 1 needs to accept less risk in order to improve performance, or reduce the determinant impacts of over-trust. As β_1 becomes negative, Agent 1 becomes increasingly risk averse which will further improve performance.

Figure 27 illustrates an instance when a risk-seeking Agent 1 learned a policy to pursue location 15. Its behavior is adjusted when Agent 2 is risk averse, where its misplaced trust is reduced. In investigating the action-value plots for each risk level, marginal improvements are also seem in state space coverage. For the risk-seeking Agent 1 at β_1=0.01, 31.46-32.23% of the state space supports trust. When Agent 1 is risk neutral, 2.16-3.78% of the state space supports trust. Ideally, the lower the share of the state space that supports trust, the more conducive for learning an independent policy.

	Outcomes- A2 "non-cooperative"			
β_1	s0s0	s0s15	s15s0	s15s15
0.0	71.04	8.12	12.78	8.06
	(1.35)	(2.35)	(1.18)	(1.88)
0.005	59.11	5.84	26.30	8.75
	(2.53)	(1.28)	(2.30)	(1.72)
0.001	70.39	7.20	14.69	7.72
	(2.51)	(1.92)	(1.52)	(1.96)
0.01	38.98	3.75	46.40	10.87
	(5.25)	(1.70)	(4.25)	(2.31)
0.1	17.82	3.72	67.12	11.33
	(4.09)	(1.47)	(3.99)	(2.60)

* reported in percent

* ten iterations of 1,000 episodes eacb

Figure 26: Final Outcomes, Interaction with "Non-cooperative" A2

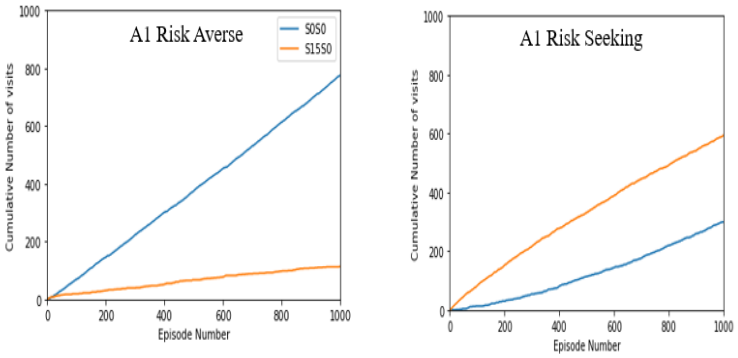

Figure 27: Learning performance, A1 Interaction with "Non-cooperative" A2

Figure 28 illustrates the third scenario where Agent 1 encounters an Agent 2 that is an "unbiased" learning agent with the same reward function. To note, Agent 2 is risk-neutral (β_2=0.0) in the scenario.

When both agents are risk-neutral, they are unable to learn to collaborate. It is not until β_1=0.01 that we see mutual collaboration rates approximately equal to mutual defection rates, at 29.39 %. Rates of over-trust outcomes are very high at 36.78 %. Increasing the risk-seeking levels only increase the over trust rates without achieving more mutual collaboration. So the ideal risk-seeking level when Agent 1 interacts with a "unbiased,risk- neutral" agent 2 is β_1=0.01.

In investigating the action-value plots, an Agent 1 at risk-seeking level of β_1=0.001 has between 6.17-7.98% of its state space favoring trust. This explains why Agent 1 is unable to obtain a policy to the interdependent reward. When Agent 1's risk-seeking level increases to β_1=0.01, the state space coverage expands to 24.54-26.49% in favor of trust, allowing for an optimal interdependent policy to be found. All in all, the scenario highlights the difficulty in obtaining high mutual collaboration

rates when the partner is independent and risk-neutral.

	Outcomes- A2 "unbiased, β_2= 0.0"			
β_1	s0s0	s0s15	s15s0	s15s15
0.0	69.48	8.81	8.44	13.27
	(1.30)	(1.26)	(2.34)	(1.19)
0.005	54.83	5.92	17.79	21.46
	(4.49)	(1.29)	(5.94)	(3.59)
0.001	68.82	7.10	9.08	15.00
	(1.49)	(2.07)	(2.40)	(2.53)
0.01	28.70	5.13	36.78	29.39
	(6.72)	(2.61)	(7.45)	(6.18)
0.1	12.80	3.56	60.72	22.92
	(2.73)	(1.15)	(2.68)	(2.88)

* reported in percent

* ten iterations of 1,000 episodes eacb

Figure 28: Final Outcomes, Interaction with "Unbiased, RN" A2

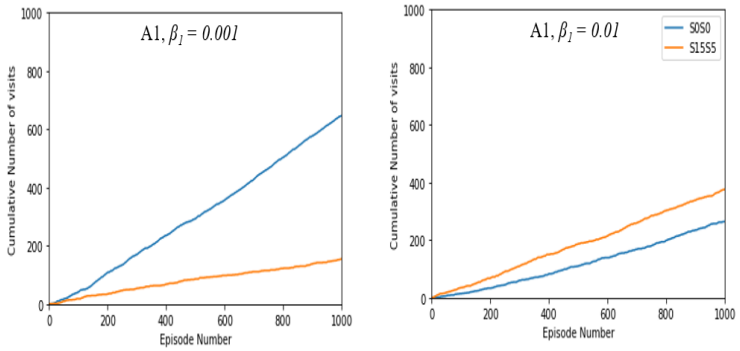

Figure 29: Learning performance, A1 Interaction with "Unbiased-RN" A2

Figure 30 illustrates the fourth scenario where Agent 1 encounters an Agent 2 that is an "unbiased" learning agent but risk-seeking (β_2=0.01). In the situation, team performance improves overall when both agents are risk-seeking at β=0.01. Now, mutual collaboration rates are much higher, and over-trust rates are much lower. Agent 1 can accept less risk if needed to β=0.001, if the detrimental impacts of the over-trust rates need to be reduced further, and still achieve majority mutual collaboration rates.

Figure 31 illustrates that under all three risk levels, there is a certain degree of mutual cooperation. Agent 1 does not learn a optimal policy towards mutual cooperation as a risk-neutral agent. When Agent 1 is risk-seeking at β_1=0.001, Agent 1 learns a optimal policy sometimes; displayed is one instance. When Agent 1 is risk-seeking at β_1=0.01, Agent 1 learns a optimal policy at all instances. Interacting with a risk-seeking Agent 2, Agent 1 learns a policy that supports trust from 6.75-7.87 %. When Agent 1 is risk-seeking at β=0.001. When Agent 1 is risk-seeking at β_1=0.01,

its policy coverage expands to 24.67-27.89 % in favor of trust.

	Outcomes- A2 "unbiased, $\beta_2 = 0.01$"			
β_1	s0s0	s0s15	s15s0	s15s15
0.0	27.82	37.60	4.35	30.23
	(6.29)	(7.25)	(1.52)	(5.20)
0.005	27.28	28.20	7.84	36.68
	(7.53)	(9.53)	(3.10)	(3.95)
0.001	27.83	36.17	5.13	30.87
	(9.82)	(9.56)	(1.46)	(5.54)
0.01	17.11	17.66	18.81	46.42
	(6.23)	(9.57)	(10.13)	(6.82)
0.1	10.62	8.18	35.62	45.58
	(3.39)	(6.08)	(10.15)	(6.89)

* reported in percent

* ten iterations of 1,000 episodes eacb

Figure 30: Final Outcomes, Interaction with "Unbiased, RS" A2

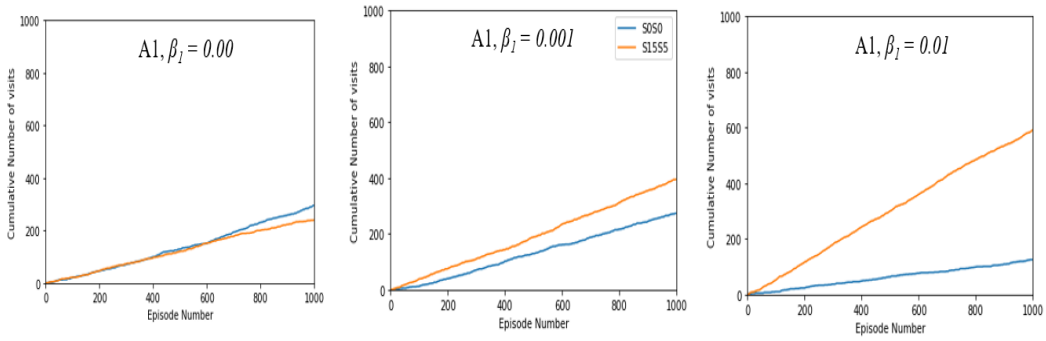

Figure 31: Learning Performance, Interaction with "Unbiased, RS" A2

6.5 Results Summary

The literature review highlighted the key characteristics of trust that come out of the psychology and sociology domain and have been mapped into computational models. However there are limitations in existing computational models that prevent us from effectively studying these characteristics. We therefore sought to create our own proposed definition and theoretical formulation to integrate these key characteristics and enable us to study these relationships. The framework enables us to study four research questions, which we investigate within a simulated environment.

The first research question seeks to determine if the proposed formulation can improve mutual collaboration. When two agents are posed with a choice between a independent and interdependent reward, the standard SARSA(λ) algorithm demonstrates that the agents will learn to act independently. It is shown that in order for the agents to learn collaboration, they are given four

capabilities. First, they are given trust as an action as a method for comparing the independent option with the interdependent option. Second, they are given current information of their partner's state which helps to reduce miscoordination events and improving mutual collaboration. Third, the agents are given the historical behavior of their partner, to calculate the risk associated with their partner. Fourth, the inclusion of the variance component in the objective function provides the capability to consider risk as part of their own decision-making process, to balance expected return with variance. There is a central role in risk to the development of trust and collaboration. This motivates the second research question.

The results of the first research question suggest that risk-seeking behavior encourage trust formation and mutual collaboration. can mutual collaboration be attained through changes in reward structure values among risk-neutral agents? Among risk-neutral agents, agents will pursue the higher expected return under the standard RL formulation. The agents will learn the optimal policies to pursue the goals of higher expected value. Mutual collaboration is achieved between the agents when the higher expected value shifts towards the interdependent reward. Correspondingly, the amount of states that value trust over not trust actions also increase. For risk-neutral agents, decisions to trust are only based on expected return. Agents need more information about their partner's behavior, in order to adjust the risk to assume in various situations. Mutual collaboration is only achieved when the interdependent reward is taken away. This further emphasizes that the agents require some level of risk-seeking behavior in order to develop trust and collaborate. This motivates the third research question.

The second research question emphasized the necessity of risk to achieve collaboration. How does varying inherent risk preferences of the agents impact trust and collaboration? These Risk-seeking behavior yields larger mutual collaboration rates, but also results in larger rates of over-trust. The results demonstrate the gap between collaboration and trust. Trust does not guarantee collaboration. High amounts of trust can lead to mis-coordinated outcomes.

The third research question identifies a trade-off. The fourth research question aims to provide insight into how agents can mitigate risks and adapt to varying partners. When Agent 1 interacts with Agent 2, some level of inherent risk-seeking behavior is required to achieve mutual collaboration. Determining the appropriate risk-seeking level for the agent requires a trade-off between mutual cooperation and over-trust or mis-coordination. Choosing the optimal risk-seeking

parameter is dependent on the partner, situation, and rewards that the agent faces. It can be seen as an optimization problem with two objectives. 1) What is the minimum amount of mutual collaboration that I want to achieve?, 2) What is the maximum amount of negative outcomes from over-trust that I am willing to accept? We must find the optimal risk-seeking parameter β that satisfies these objectives in the current situation.

Figure 32 summarizes the final performance of the developed SARSA-Trust algorithm, tested on a scenario where the rewards are $R_{ind} = 9.5$, $R_{p,int} = 20$, and, $R_{n,int} = -1$. With the standard SARSA algorithm, the agents are unable to learn a policy to mutual collaborate. Instead, both agents learn to act independently 84.85 % of the time. With the SARSA-Trust algorithm and both agents set as risk-seeking at β=0.10, mutual cooperation rates of 62.54 % is achieved.

	Treatment effects			
	s0s0	s0s15	s15s0	s15s15
Tests				
SARSA(λ)	84.85	6.60	7.80	0.68
(sd)	(1.23)	(1.05)	(0.85)	(1.75)
SARSA-T, β = 0.00	70.02	7.68	8.08	14.22
(sd)	(1.79)	(2.32)	(2.43)	(2.67)
SARSA-T, β = 0.01	23.06	13.44	19.61	43.89
(sd)	(5.84)	(9.03)	(7.65)	(4.49)
SARSA-T, β = 0.10	6.24	14.51	16.71	62.54
(sd)	(4.61)	(6.71)	(8.18)	(4.12)

Figure 32: Final Outcomes Summary

7 Conclusions

7.1 Conclusions

Multi-agent systems prolific in society. The relationships and interconnectedness between various autonomous systems and humans are new challenges. Working together as a team is often the goal, but is very challenging to achieve. Trust is the primary determinant of collaborative outcomes among multi-agent systems. The research seeks to investigate the role of reward, risk, and behavior in trust formation and collaboration among two agents. We seek to integrate the following proposed definition of trust into a reinforcement learning framework; trust is the willingness to take on the risk governed by the behavior of another.

Trust has been extensively studied within psychology and sociology. The work of Castelfranchi and Falcone map significant characteristics of trust into the computational domain [7]. Wagner illustrates the importance to consider interdependency into a trust formulation as a decision process [18], [19]. Contemporary computational models do not allow us to adequately investigate the characteristics of interdependency, risk, and behavior in the context of learned behavior. Moreover, existing computational trust models built within learning frameworks do not meaningfully express trust in terms of accepted definitions.

We start by proposing our definition of trust as the extent to which an agent is willing to take on risk (the variance in reward) governed by the behavior of another agent. The proposed definition is consonant with accepted definitions from the domains of psychology and sociology. The proposed definition is used to create a theoretical framework using the key characteristics of trust illustrated by the work of Castelfranchi and Falcone, and Wagner.

The RL architecture to integrate the two theories. Trust is formulated as a decision between an independent and interdependent goal. Additionally, the objective function includes a variance component. This drives the agent's decision-making process to balance maximizing expected return with the risk (variance) in outcomes governed by the partner. The theoretical framework enables us to investigate under what conditions do agents collaborate, and what does the value of trust reveal about the nature of these relationships. Specifically, the formulation allows us to investigate four research questions concerning components of the formulation, reward values, risk preferences, and partner behavior. We investigate these research questions within a gridworld simulated environment.

Two independent learning agents are given the standard SARSA(λ) algorithm and given the choice between an independent and interdependent reward. The SARSA(λ) algorithm is unable to encourage the agents to collaborate. The theoretical framework, and the corresponding SARSA-Trust algorithm enables the agents to achieve mutual collaboration rates of 62.54 %. The agents required four key components to enable collaboration. First, they are given trust as an action as a method for comparing the independent option with the interdependent option. Second, they are given current information of their partner's state which helps to reduce miscoordination events and improving mutual collaboration. Third, the agents are given the historical behavior of their partner, to calculate the risk associated with their partner. Fourth, the inclusion of the variance

component in the objective function provides the capability to consider risk as part of their own decision-making process, to balance expected return with variance.

The second research question revealed the risk-neutral agents were unable to collaborate unless the independent reward was removed. This suggests the central role risk plays in trust formation and collaboration. The third research question identified that increases in risk-seeking behavior results in increases in mutual collaboration. However, with increases in risk-seeking behavior, there is an increase in mis-coordination rates, specifically, over-trust rates, instances where one agent trusts the partner to its own detriment. In the fourth research question, it is shown that agents can adjust their risk preferences to mitigate the detrimental effects of over-trust outcomes. Agents can adapt to their situation and their partner to achieve the optimal cooperation levels.

7.2 Contributions

1. This research has created a comprehensive definition of trust that integrates six important characteristics together within a theoretical framework.

2. The theoretical framework enables the study of trust and collaboration with a more nuanced characterization of reward, risk, and behavior.

3. Within a simulated environment, results illustrate that trust formation and collaboration are not attainable within risk-seeking behavior. There is a trade-off between risk, trust, and collaboration. Agents can mitigate risk by adapting to situations to address these trade-offs.

4. Targeted publication venues: IEEE Transactions on Human-Machine Systems, ACM Transactions on Human-Robot Interaction

7.3 Limitations and Future Work

The results highlight three major limitations in the current work.

The primary limitation is in the variance formulation. The agent requires a way to be informed about the uncertainty from the interdependent reward at a particular state. In the current framework, there are many states that are not visited enough. Therefore, the incremental expected return and incremental variance calculated through reinforcement learning methods are very small. We choose to calculate the variance of the history of interdependent rewards, rather than a state-specific calculation of variance. Extensions of this work can consider how to improve the calculation

of uncertainty of the expected return from the interdependent reward at a particular state-action pair. This could be achieved through safe RL methods that influence the agent's exploration process. A better calculation of risk can articulate trade-offs between risk and states. Stochastic policy gradient methods can be explored. Safe RL literature also suggests exploration methods for risk mitigation as a viable alternative.

The agent using this approach is also limited in adaptability. Optimal risk levels for a given partner are determined through a manual process. Ideally, the agent should determine the ideal parameter by itself, and adapt when the partner's behavior changes.

Future work can focus on validating the framework against existing trust models and real data. Furthermore, the approach considers a simple interdependency situation. Interdependency theory offers a lot more insight to expand interdependent situations. we need to know the rewards.

The approach considers the interaction of two agents. How would this approach be adopted in a n-size multi-agent system. Additionally, trust is formed through experiences. Other MAS models use other sources of information to inform trust in a larger network.

References

[1] Michael Wooldridge, *An Introduction to MultiAgent Systems*, 2nd. Cambridge, MA, USA: John Wiley and Sons, 2009, ISBN: 0470519460.

[2] Michael C. Riddell, Dessi P. Zaharieva, Loren Yavelberg, Ali Cinar, Veronica K. Jamnik, "Exercise and the development of the artificial pancreas: One of the more difficult series of hurdles," *Journal of Diabetes Science and Technology*, vol. 9(6), pp. 1217–1226, 2015.

[3] R. Burbaite, V. Stuikys and R. Damasevicius, "Educational robots as collaborative learning objects for teaching computer science," in *2013 International Conference on System Science and Engineering (ICSSE)*, 2013, pp. 211–216.

[4] Prashant P. Reddy and Manuela M. Veloso, "Strategy learning for autonomous agents in smart grid markets," *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, IJCAI'11, pp. 1446–1451, 2011.

[5] Christopher Prinz, Friedrich Morlock, Sebastian Freith, Niklas Kreggenfeld, Dieter Kreimeier, Bernd Kuhlenkötter, "Learning factory modules for smart factories in industrie 4.0," *Procedia CIRP*, vol. 54, pp. 113–118, 2016, 6th CIRP Conference on Learning Factories, ISSN: 2212-8271.

[6] Denise Rousseau, Sim Sitkin, Ronald Burt, and Colin Camerer, "Introduction to special topic forum: Not so different after all: A cross-discipline view of trust," *The Academy of Management Review*, vol. 23, no. 3, pp. 393–404, 1998.

[7] Cristiano Castelfranchi and Rino Falcone, *Trust Theory: A Socio-Cognitive and Computational Model*. John Wiley and Sons, Ltd, 2010.

[8] Harold Kelley, Caryl Rusbult, Harry Reis, John G. Holmes, Norbert L. Kerr, and Paul A. M. Lange, *An Atlas of Interpersonal Situations*. Cambridge University Press, 2003.

[9] Roger C. Mayer, James H. Davis and F. David Schoorman, "An integrative model of organizational trust," *Academy of Management Review*, vol. 20, pp. 709–734, 1995.

[10] Jin-Hee Cho, Kevin Chan, and Sibel Adali, "A survey on trust modeling," *ACM Computer Survey*, vol. 48, no. 2, p. 40, 2015.

[11] Wanita Sherchan, Surya Nepal, and Cecile Paris, "A survey of trust in social networks," *ACM Computer Survey*, vol. 45, no. 4, p. 33, 2013.

[12] Audun Josang and Roslan Ismail, "The beta reputation system," in *Bled eConference*, 2002.

[13] Jordi Sabater and Carles Sierra, "Reputation and social network analysis in multi-agent systems," in *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, ser. AAMAS '02, Bologna, Italy: ACM, 2002, pp. 475–482, ISBN: 1-58113-480-0.

[14] Gehao Lu, Joan Lu, Shaowen Yao, and Jim Yip, "A review on computational trust models for multi-agent systems," *The Open Information Science Journal*, vol. 2, pp. 18–25, 2009.

[15] Trung Dong Huynh, Nicholas R. Jennings, and Nigel R. Shadbolt, "An integrated trust and reputation model for open multi-agent systems," *Autonomous Agents and Multi-Agent Systems*, vol. 13, no. 2, pp. 119–154, 2006.

[16] W.T. Luke Teacy, Jigar Patel, Nicholas R. Jennings, and Michael Luck, "Travos: Trust and reputation in the context of inaccurate information sources," *Autonomous Agents and Multi-Agent Systems*, vol. 12, no. 2, pp. 183–198, 2006.

[17] Kevin Regan, Pascal Poupart, Robin Cohen, "Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change," in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, ser. AAAI'06, Boston, Massachusetts: AAAI Press, 2006, pp. 1206–1212, ISBN: 978-1-57735-281-5.

[18] Harold Kelley and Thibaut John W., *Interpersonal Relations: A Theory of Interdependence*. John Wiley and Sons, 1978.

[19] Alan Wagner and Paul Robinette, "Towards robots that trust: Human subject validation of the situational conditions for trust," *Interaction Studies*, vol. 16, no. 1, pp. 89–117, 2015.

[20] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa, "Planning with trust for human-robot collaboration," *CoRR*, vol. abs/1801.04099, 2018. arXiv: 1801.04099.

[21] Richard S. Sutton and Andrew G. Barto, *Introduction to Reinforcement Learning*, 1st. Cambridge, MA, USA: MIT Press, 1998, ISBN: 0262193981.

[22] Thomas Tran and Robin Cohen, "Improving user satisfaction in agent-based electronic marketplaces by reputation modelling and adjustable product quality," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, ser. AAMAS '04, New York, New York: IEEE Computer Society, 2004, pp. 828–835, ISBN: 1-58113-864-4.

[23] Li Haiyan, "Dynamic trust game model between venture capitalists and entrepreneurs based on reinforcement learning theory," *Cluster Computing*, pp. 1–12, 2018.

[24] Abdullah Aref and Thomas Tran, "A hybrid trust model using reinforcement learning and fuzzy logic," *Computational Intelligence*, vol. 34, pp. 515–541, 2018.

[25] Kevin Regan and Robin Cohen, "A model of indirect reputation assessment for adaptive buying agents in electronic markets," 2005.

[26] Javier Garcia and Fernando Fernandez, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, pp. 1437–1480, 2015.

[27] Matthias Heger, "Consideration of risk in reinforcement learning," *Proceedings of the 11th International Conference on Machine Learning*, pp. 105–111, 1994.

[28] Abhijit Gosavi, "Reinforcement learning for model building and variance-penalized control," *Proceedings of the Winter Simulation Conference*, pp. 373–379, 2009.

[29] Peter Geibel and Fritz Wysotzki, "Risk-sensitive reinforcement learning applied to control under constraints," *Journal of Artificial Intelligence Research,*, vol. 24, pp. 81–108, 2005.

[30] Vivek S. Borkar, "Q-learning for risk-sensitive control," *Mathematics of Operations Research*, vol. 27(2), pp. 294–311, 2002.

[31] Clement Gehring and Doina Precup, "Smart exploration in reinforcement learning using absolute temporal difference errors," *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1037–1044, 2013.

[32] Charles A. Holt, *Markets, Games, and Strategic Behavior: An Introduction to Experimental Economics (Second Edition)*. Princeton University Press, 2019.

[33] Harry Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7(1), pp. 77–91, 1952.