Metastrategies in the Colored Trails Game

Andreas ten Pas
B.Sc. Thesis
Department of Knowledge Engineering
Maastricht University

August 2010

Abstract

This article investigates the mapping of Colored Trails onto existing games. Two metastrategies are introduced to analyze the game. Based on those metastrategies, a two-player normal form game is derived. Depending on parameter settings, this game can be mapped onto two well-studied games from the field of game theory, i.e. Prisoner's Dilemma and Stag Hunt. By introducing a third metastrategy, a three-strategy version evolves. It turns out that for an arbitrary number of metastrategies, Colored Trails falls apart into a predetermined number of different games which are similar to Prisoner's Dilemma and Stag Hunt. Experiments to train learning automata to play the two- and the three-strategy games are performed. Moreover, the impact of inequity aversion on those games is examined. Learning automata without inequity aversion converge to Nash equilibria in those games, while automata with inequity aversion rapidly converge to the strategy with the highest joint payoff, i.e. the Pareto-optimal outcome.

1 Introduction

Research in the field of multi-agent systems is recently following a trend which departs from purely rational and self-interested agents [6]. Traditionally, agents were designed based on the principles of classical game theory. However, those principles do not correspond perfectly to the behavior of humans, as research in the field of Behavioral and Welfare Economics has shown (see e.g. Fehr and Schmidt [4] or Chevaleyre et al. [2]).

The games under study become more and more complex. One game that is currently receiving a lot of attention from researchers is Colored Trails (CT), a testbed for multi-agent systems in which agents exchange chips with each other to achieve predefined goals. Previous studies of CT mainly involved the design of agent architectures and the investigation of human behaviour (see e.g. Glaim [7] or Hennes [8]). Therefore, in this study, we follow a more general approach to analyze the game by means of classical game theory.

For many games, the behavior of humans has been studied and captured into descriptive models (see e.g. Fehr and Schmidt [4]). One such model that represents the preference of humans to be fair and to minimize inequitable outcomes, is inequity aversion. We also apply this concept to Colored Trails.

1.1 Research questions

Due to the complexity of the Colored Trails game, concrete strategies that allow optimal play are hard to find directly. To better analyze the game, metastrategies could become a significant tool. In this study, we therefore focus on the following problem statement:

How can meaningful metastrategies for Colored Trails be established?

This statement leads to the following research questions:

1. Which abstractions or generalizations need to be made?
2. Onto which games can Colored Trails be mapped?

For many existing games, inequity aversion is an interesting concept [3]. We also investigate this concept in Colored Trails, leading to the following research question:

3. What is the role of inequity aversion in Colored Trails?

1.2 Structure

This article is structured as follows. Section 2 gives the required background in game theory and introduces the Colored Trails game. Section 3 discusses learning automata which allow us to study learning dynamics in the investigated games. Inequity aversion is presented in Section 4. Then, we give our main contributions. In Section 5, we derive metastrategies for Colored Trails. Section 6 reports the experiments and discusses their results. Conclusions are given in Section 7.

2 Background

Game Theory studies the interaction of agents in strategic situations where the success of the action of one agent depends on the actions of all other agents. While game theory has been extensively applied in economics, biology and the political and social sciences, it also contributes to multi-agent systems.

In game theory, strategic situations are modeled in the form of games. The following three components compose a game: players, actions and payoffs. By choosing actions, players make decisions, and for each combination of their actions, there is a payoff for each of them.

In the following subsections, the concept of games in normal form, the solution concepts used in game theory and two examples of games in normal form, which are relevant for this study, are presented, i.e. Prisoner's Dilemma and Stag Hunt. For more details on Game Theory, we refer to Leyton-Brown and Shoham [10].

2.1 Games in Normal Form

The most familiar representation of strategic interactions in game theory is the normal form game, also known as the strategic form game. The following definition is from Leyton-Brown and Shoham [10].

Definition 1. *A finite, n-person normal form game is a tuple (N, A, U) where:*

1. $N = 1, \dots, n$ *is a finite set of players, indexed by i,*
2. $A = A_1 \times \cdots \times A_n$*, where* A_i *is the set of actions available to player i, and*
3. $U = u_1, \dots, u_n$*, where* $u_i : A \rightarrow \Re$ *is a real-valued payoff function for player i.*

In normal form games, players interact simultaneously. Each player i selects an action a_i from its action set A_i. Then, the payoff for each player is given by the joint action $a = a_1, \dots, a_n \in A$.

To specify which action a player takes in each situation, a strategy π_i is used. The strategy π_i is called a pure strategy if each situation is mapped to a single action, while it is called a mixed strategy if the actions are chosen according to a probability distribution. To specify a strategy π_i, a single probability distribution can be used. Then, π_{ij} is the probability of player i to take action a_j from its action set A_j, and $\sum_j^n \pi_{ij} = 1$. The assignment of a strategy to each player is the strategy profile $\pi = \{\pi_1, \dots, \pi_n\}$.

A normal form game with two players can be represented as a matrix in which each row corresponds to a possible action for player 1 and each column corresponds to a possible action for player 2. The cells of the matrix give the payoffs for the two players. First listed is player 1's payoff, followed by the payoff for player 2.

2.2 Solution Concepts

The main question in strategic interactions is: "What should I do?". To answer this question, game theory provides us with a number of solution concepts which evaluate the strategy profile π. Relevant for this study are the concepts of best response, dominant strategies, Nash equilibrium and Pareto optimality. The following definitions are from Wooldridge [13].

Definition 2. *A strategy π_i is player i's best response to a strategy π_j by player j if it gives the highest payoff when played against π_j.*

Definition 3. *A strategy π_i is dominant for player i if it is the best response to all of player j's strategies.*

This definition means that, no matter what strategy π_j player j chooses, player i will profit at least as much from playing π_i than it would do from anything else.

Definition 4. *Players are in Nash Equilibrium if, given that the other players remain at their strategies, no player can do better by changing its strategy.*

In other words, player i and player j which play strategies π_i and π_j, respectively, are in Nash equilibrium if, under the assumption that player j plays π_j, player i can make no better choice than play π_i, and, under the assumption that player i plays π_i, player j can make no better choice than play π_j.

Definition 5. *A strategy π_i is Pareto optimal if no player can improve his payoff by changing its strategy without making another player worse off.*

2.3 Prisoner's Dilemma

The Prisoner's Dilemma (PD) has been popularized by Axelrod in 1984 [1]. Its story is as follows: Two suspects are arrested for a crime. They are taken to separate interrogation rooms, and each suspect can either confess to the crime (cooperate) or deny it (defect). If they both confess, they go to prison for a year. If one suspect denies, i.e. he supplies some evidence that incriminates himself, then that suspect is freed, and the other one is imprisoned for nine years. If both deny, then they are imprisoned for six years. The payoff matrix for the original PD is given in Table 1.

	confess	deny
confess	-1,-1	-9,0
deny	0,-9	-6,-6

Table 1: Payoff matrix for the Prisoner's Dilemma.

In this game, the dominant strategy for both players is to deny. The reason for this is that each player always gets a higher expected payoff if it plays deny, no matter what the opponent does. Therefore, the only Nash equilibrium here is *(deny, deny)*, while *(cooperate, cooperate)* remains as the Pareto optimal solution.

The generalized form of the PD, in which positive payoffs are used, is given in Table 2.

	cooperate	defect
cooperate	R,R	S,T
defect	T,S	P,P

Table 2: Canoncial form of the payoff matrix for the Prisoner's Dilemma ($T > R > P > S$).

The constants used in the table have the following meanings: T for Temptation to defect, S for Sucker's payoff, P for Punishment for mutual defection and R for Reward for mutual cooperation.

To define a game as PD, the following inequalities must hold [9]:

$$T > R > P > S. \tag{1}$$

This condition ensures that the only Nash equilibrium in the PD is for both players to defect and that it is Pareto optimal to cooperate for both players.

2.4 Stag Hunt

The Stag Hunt (SH) is a game that is based on a story by the French philosopher, Jean Jacques Rousseau, that tells the following situation [12]: at the same time, two hunters go out to acquire food; they can hunt either for a stag (cooperate) or for rabbits (defect). Hunting stags is difficult and requires that both hunters cooperate, but a stag provides a lot of meat. A rabbit will provide less meat, but each hunter can catch one easily. If one of them hunts a stag alone, the chance of capturing one is minimal.

The game that accompanies this story is given in Table 3.[1]

	stag	rabbit
stag	10,10	0,8
rabbit	8,0	4,4

Table 3: The payoff matrix for the Stag Hunt.

In this game, there are two Nash equilibria, one at *(stag, stag)* and the other one at *(rabbit, rabbit)*. The former is Pareto optimal, but the latter is less risky.

The general form of the SH is given in Table 4.[2] Here, the following inequalities must hold:

$$R > T \geq P > S. \tag{2}$$

This conditions ensures that there are two Nash equilibria in the game and that one of them is Pareto optimal, while for the other, there is less risk involved for both players.

	stag	rabbit
stag	R,R	S,T
rabbit	T,S	P,P

Table 4: Canonical form of the payoff matrix for the Stag Hunt ($R > T \geq P > S$).

2.5 Colored Trails

Colored Trails (CT) is a framework to study cooperation in multi-agent systems, developed at Harvard University, School of Engineering and Applied Sciences [5]. CT is a board game played on a board of $m \times n$ squares each colored in one of k colors. One or more squares on the board can be assigned to be goal states. Each player has a set of colored chips and a piece located on the game board. A player can use a colored chip to move his piece to an adjacent square (left, right, up and down) of the same color as the chip. A goal state is given to each player and multiple players may share the same goal. The game is played in cycles of two consecutive phases: communication and movement. During the first phase, the players are allowed to interchange chips with each other, while during the second phase, they can move their pieces on the board. Players can do multiple moves in the second phase.

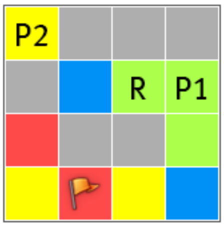


Figure 1: A possible board configuration in Colored Trails.

Game Configuration In this study, a three-player negotiation variant of CT is used [8]. The set of players contains two proposers and one responder. Proposers can propose a chip exchange to the responder. The responder can accept either one proposal or none at all. All players know the board state. The responder can see all chip sets, while proposers only have knowledge about their own chip set and the one of the responder. This variation of CT is played as a one-shot game. Proposers can only offer a single proposal and the responder can only accept or reject. Once the responder reacted, the chips are interchanged according to the winning

[1] One could argue that hunting rabbits together should give the same payoff to a player as hunting rabbits on its own. However, in this case, it is assumed that, e.g., there are two rabbits and thus one hunter can easily capture both of them, if the other hunter is not trying to get any; thus resulting in a higher payoff. Otherwise, the two hunters need to divide the rabbits among each other.

[2] We restrict ourselves to a symmetrical version of the game, similar to the PD.

proposal or stay fixed if the responder rejected both offers. Then, the best possible sequence of moves is automatically computed and each player receives a personal score. Here, the following scoring function is used:

$$s = 100g + 10c - 25d \tag{3}$$

where $g \in \{0, 1\}$ represents whether the player reached the goal ($g = 1$) or did not reach it ($g = 0$), c is the number of chips the player has left and d is the distance to the goal.

3 Learning automata

Stochastic learning automata (SLA) are a class of automata. An automaton is a computational model of a complex system, e.g. an agent playing a game. The term stochastic refers to the ability of the automaton to adapt to changes in its environment. This ability is the result of the learning process performed by the automaton where learning is defined as any particular change in the behavior of the automaton based on experience.

SLA learn what to do without any available information on the optimal action. The automaton randomly selects one action and updates its action probabilities depending on the response it obtained from the environment, e.g. according to the payoff matrix of a game. This process is repeated until a certain goal is achieved or a certain number of iterations is reached.

A particular class of SLA are finite-action set learning automata (FALA). FALA are assumed to be in a stateless environment in which the utility of the current action is independent of previous actions performed by the automaton or other agents.

Definition 6. *A finite-action set learning automaton is a tuple* (α, β, π) *where:*

- $\alpha = \{1, \ldots, N\}$ *is a finite set of actions the automaton can choose from,*
- $\beta \in (0, 1)$ *is a set of environment responses and*
- $\pi = \pi_1, \ldots, \pi_n$ *is a set of probabilities over* α*, i.e.* π_i *is the probability to choose action* a_i*.*

To update the action probabilities π, several learning schemes can be used. Linear schemes update probability π_i at iteration $(t + 1)$ based on the response $\beta_i(t)$ obtained by performing action a_i at the current iteration t. The following linear scheme, called reward-inaction scheme, has been shown to converge to equilibria in games [11]:

$$\pi_i(t+1) = \begin{cases} \pi_i(t) + \lambda\beta_i(t)(1 - \pi_i(t)) & \text{if } i = j, \\ \pi_i(t) - \lambda\beta_i(t)\pi_i(t) & \text{otherwise.} \end{cases} \tag{4}$$

Here, $\lambda \in (0, 1)$ is the learning rate or step-size associated to reward response from the environment. Equation 4 gives a probability distribution if $\beta(t)$ is continuous and $\beta(t) \in [0, 1]$.

FALA can be situated in a game to represent players learning to play optimal strategies. At iteration t, a collection of n agents select their actions $a_1, \ldots, a_n$ according to the strategy profile π. The environment's response $\beta_i(t)$ for agent i, or automaton i, is the same as the reward or the utility u_i obtained by performing the joint action $a_1, \ldots, a_n \in A$ for player i.

Since Equation 4 requires $u(t) \in [0, 1]$, the payoffs of the game need to be normalized.

4 Inequity Aversion

To explain human behavior in games, the fields of Behavioral and Welfare Economics provide us with a number of different models of *fairness*. One of those models is inequity aversion (IA), developed by Fehr and Schmidt [4].

IA assumes that there are not only players that are purely rational and selfish, but also players who are unhappy with inequitable rewards. Here, a reward is regarded as inequitable by a player if there are other players that are better off or other players that are worse off. The utility function that accompanies IA is denoted by

$$\begin{aligned} u_i = x_i &- \frac{\alpha_i}{n-1} \sum_{j \neq i} \max[x_j - x_i, 0] \\ &- \frac{\beta_i}{n-1} \sum_{j \neq i} \max[x_i - x_j, 0], \end{aligned} \tag{5}$$

where x_i is the reward for the current player i, x_j is the reward for player j, and α and β are parameters weighting different forms of inequity. It is assumed that $\beta_i \leq \alpha_i$ and $0 \leq \beta_i < 1$. The second and third term in the above equation weigh the utility other players loose against the utility they gain. For the two-player case, Equation 5 simplifies to

$$u_i = x_i - \alpha \max[x_j - x_i, 0] - \beta \max[x_i - x_j, 0]. \tag{6}$$

5 Metastrategies in Colored Trails

The CT game, and its three-player negotiation variant, is receiving increasing interest. Here, we analyze whether the game has certain concepts in common with existing and well-studied games.

A large space of possible initial game situations arises from the combination of different board configurations and chip sets assigned to the players. Therefore, in this study, we introduce metastrategies to provide a better analysis of the CT game based on the following sensible abstractions.

1. Proposers and responders do not hurt themselves, i.e. they are not taking actions that could decrease their scores in the game.

2. The responder plays a static strategy. If there is any best proposal, he always accepts it. If both proposals are equally good, he randomly accepts one of them. The responder regards all proposals as acceptable which are not reducing his own score (see Abstraction 1). Other proposals are always rejected. This abstraction ensures that the proposer needs to offer a deal which helps the responder since other proposals are simply rejected by the responder.

Given this static strategy, the game reduces to a two-player competition between the proposers. For proposers, we identify the following two extreme strategies: focus on its own gain (*H*) or focus on the responder's gain (*L*). The first strategy increases the proposer's chance of getting the highest score in the game, but the other proposer can prevent this by offering *L*, a deal that the responder prefers. The second strategy increases the proposer's chance of having the deal that is accepted by the responder, but the expected gain for the proposer is low. These strategies lead us to the next abstraction:

3. In a two-strategy game, we only consider the extreme strategies *H* and *L*.

Intermediate strategies will be studied later in this section.

5.1 Two Metastrategies

We are left with a set of two metastrategies a proposer can choose from: it can either not help the responder and thereby get a high increase to its own score (*H*), or it can help the responder and thereby only get a low increase to its own score (*L*).

In this way, the reduced game fits to the form of a two-player normal form game with two actions (which are equivalent to the metastrategies) and can be represented by the matrix in Table 5.

	H	L
H	$\frac{1}{2}A, \frac{1}{2}A$	$0, B$
L	$B, 0$	$\frac{1}{2}B, \frac{1}{2}B$

Table 5: Payoff matrix for the reduced two-player Colored Trails game.

Given no exchange of chips, the proposers achieve a certain score that is based on the initial board configuration and the initial chip sets. The game in Table 5 models the gain of a chip exchange with the responder. Here, A is the gain for the better deal for the proposer, i.e. not helping the responder, and B is the gain for the worse deal for the proposer, i.e. helping the responder. The cells of the matrix represent the following game situations:

1. *(H,H)*: Both players propose deals which would not help the responder but raise their own reward, thereby trying to achieve score A. To the responder, it does not matter which deal to choose, thus it accepts each of them with equal probability. Therefore, the expected gain for each proposer is $\frac{1}{2}A$.
2. *(H,L)* and *(L,H)*: One player proposes a deal that makes the responder better off, while the other proposer's deal makes itself better off. The deal that makes the responder better off is accepted. The proposer who offered this deal gains B, while the other proposer does not gain anything.
3. *(L,L)*: Both players propose deals which would help the responder, thereby trying to achieve score B. Since both proposals are equally good for the responder, it again accepts each of them with equal probability. Therefore, the expected reward for each proposer is $\frac{1}{2}B$.

By changing the constants in Table 5, two kinds of typical normal form games can originate.[3] If $A > 2B$, the game can be classified as Stag Hunt. If $A < 2B$, the game can be classified as Prisoner's Dilemma. Examples of the two games which fit those requirements are given in Tables 6 and 7.

	H	L
H	$3, 3$	$0, 2$
L	$2, 0$	$1, 1$

Table 6: Example payoff matrix for the Stag Hunt version of the two-player Colored Trails game, with $A = 6$ and $B = 2$.

	H	L
H	$3, 3$	$0, 4$
L	$4, 0$	$2, 2$

Table 7: Example payoff matrix for the Prisoner's Dilemma version of the two-player Colored Trails game, with $A = 6$ and $B = 4$.

5.2 Three Metastrategies

Looking back at the CT Game, the two metastrategies of helping the responder or not helping it, might not cover all options the proposers have because they are the most *extreme* options. Therefore, we introduce a third, intermediate metastrategy, M, which refers to a deal that is higher than the low deal, L, but lower than the high deal, H. The two-player, three-strategy game that includes this new strategy is given in Table 8 where strategy H gives payoff A, strategy M gives payoff B and strategy L gives payoff C to the proposer. By definition, $A > B > C$.

The cells of the matrix in Table 8 represent the following game situations:

[3] The special case of $A = 2B$ is not regarded in this study.

	H	M	L
H	$\frac{1}{2}A, \frac{1}{2}A$	$0, B$	0,C
M	$B, 0$	$\frac{1}{2}B, \frac{1}{2}B$	0,C
L	$C, 0$	$C, 0$	$\frac{1}{2}C, \frac{1}{2}C$

Table 8: Payoff matrix for the reduced two-player, three-strategy Colored Trails game.

- *(H,H)*, *(M,M)* and *(L,L)*: Both players propose equal deals. To the responder, it does not matter which deal to choose, thus both deals are accepted with equal probability. The expected gain for both proposers is $\frac{1}{2}$ times their wanted reward.
- *(H,M)*, *(H,L)*, *(M,H)*, *(M,L)*, *(L,H)*, *(L,M)*: Both players propose different deals. The deal that makes the responder better off is accepted. The proposer who offered that deal gains its wanted reward, while the other proposer does not gain anything.

Independent of the parameters in the matrix in Table 8, there is always a Nash equilibrium at *(L,L)* because playing L is the best response strategy for both players. Two other game situations, i.e. *(M,M)* and *(H,H)*, could become Nash equilibria, depending on the parameters of the game. If $A > 2B$, then *(H,H)* becomes a Nash equilibrium. If $B > 2C$, then *(M,M)* becomes a Nash equilibrium. Therefore, if $A > 2B > 4C$, all three strategies where both players choose the same action become Nash equilibria.

This analysis provides us with the following four games for which, for convenience, the same names as for the two-player, two-strategy games are used:

1. A SH with three equilibria at *(H,H)*, *(M,M)* and *(L,L)*, if $A > 2B > 4C$.
2. A combination of PD and SH with two equilibria at *(M,M)* and *(L,L)*, if $A < 2B$ and $B > 2C$.
3. A SH with two equilibria at *(H,H)* and *(L,L)*, if $A > 2B$ and $B < 2C$.
4. A PD with one equilibrium at *(L,L)*, if $A < 2B$ and $B < 2C$.

	s_1	s_2	. . .	s_n
s_1	$\frac{1}{2}r_1, \frac{1}{2}r_1$	$0, r_2$	. . .	$0, r_n$
s_2	$r_2, 0$	$\frac{1}{2}r_2, \frac{1}{2}r_2$	. . .	$0, r_n$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
s_n	$r_n, 0$	$r_n, 0$	. . .	$\frac{1}{2}r_n, \frac{1}{2}r_n$

Table 9: Payoff matrix for the reduced two-player, n-strategy Colored Trails game.

5.3 Generalized Metastrategies

In general, given that each player has a set of n metastrategies, CT falls apart into 2^n different games. Each of those games can then be identified as an extension of a SH, a PD or a combination of both. The general structure of such a two-player, n-strategy game is given in Table 9, where s_i is the i-th strategy of each player and r_i is the reward associated with that strategy. Given strategies s_i and s_{i+1}, it is assumed that, for the proposer, the reward for strategy s_i is higher than the reward for strategy s_{i+1}, and for the responder, the reward for strategy s_i is lower than the reward for strategy s_{i+1}. If the proposers offer two different deals, the responder always accepts the deal that makes itself better off and rejects the other. This gives the wanted reward to the player whose proposal is accepted and 0 to the other player. If both proposers offer the same deal, the responder accepts each of them with equal probability and thus the expected gain is $\frac{1}{2}$ times the wanted reward of the proposer.

The Nash equilibria in the two-player, n-strategy game can be found on the diagonal of its payoff matrix. There are no equilibria outside the diagonal and there is always an equilibrium at *(s_n,s_n)*. The number of potential Nash equilibria that can be found in the game is n.

6 Experiments and Results

In this section, we first introduce the tools to analyze the dynamics of learning automata in games, then we specify the parameters and initial conditions of the experiments and give results. Finally, those results are discussed.

6.1 Methodology

To examine the learning dynamics of FALA in the reduced CT games and to compare between learning with inequity aversion and without, we present two visual methods to perform this analysis, i.e. policy trajectory plots and direction field plots.

Policy Trajectory Plots The evolvement of strategies in a two-player, two-strategy normal form game can be plotted by means of a trajectory plot. Since the probabilities for the actions of each player add up to one, the probabilities for the second action of both players can be calculated as: $\pi_{12} = 1 - \pi_{11}$ and $\pi_{22} = 1 - \pi_{21}$. Therefore the strategy profile $\pi = \{\pi_1, \pi_2\}$ can be reduced to the pair (π_{11}, π_{21}) without losing information. During one single or multiple runs, the trajectory of this pair is recorded and plotted in a two-dimensional space. Grayscales are used to illustrate the direction and speed of convergence. The higher the number of iterations of the learning algorithm is at, the darker is the color of the trajectory.

For the two-player, three-strategy game, the strategy profile $\pi = \{\pi_1, \pi_2\}$ cannot just be reduced to be plotted in a

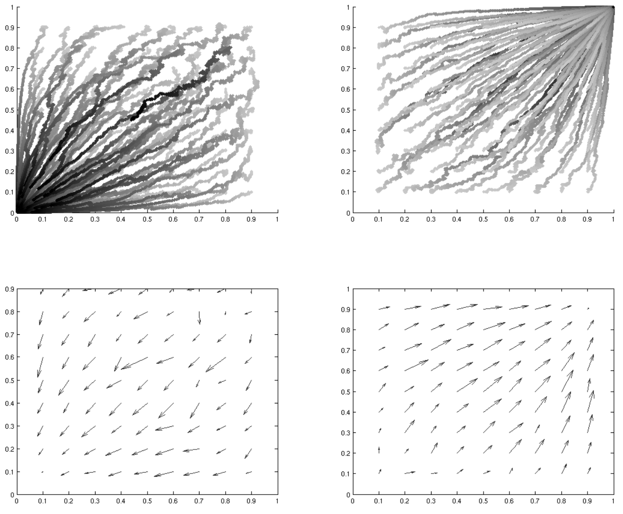


Figure 2: Trajectory plots of FALA in the Prisoner's Dilemma. The plots in the left column display trajectories of automata without inequity aversion, while the plots in the right column display those with inequity aversion.

two-dimensional space. However, the trajectory of each strategy π_i can be displayed in a single ternary plot by recording the evolvement of the tuple $(\pi_{i1}, \pi_{i2}, \pi_{i3})$. The vertices of the triangle in this plot correspond to the pure strategies $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$.

Direction Field Plots Another way to illustrate the evolvement of strategies in a two-player, two-strategy normal form game are direction field plots in which arrows indicate the direction and velocity of movement of the strategies through the probability space. The arrows start at regular grid points over $[0, 1]^2$. For grid points $\pi_{11}(t_0), \pi_{21}(t_0) \in [0, 1] \times [0, 1]$, the velocity field is then given by

$$\frac{d(v,u)}{dt} = \frac{(\pi_{11}(t_0 + \Delta t) - \pi_{11}(t_0), \pi_{21}(t_0 + \Delta t) - \pi_{21}(t_0))}{\Delta t}, \tag{7}$$

where Δt is the number of iterations spent at each grid point and v and u represent the strategies of the first player and the second player, respectively. The arrows based on Equation 7 point in the direction of $\frac{d(v,u)}{dt}$.

6.2 Experimental Setup

In the following experiments, for the two-strategy games, the maximum number of iterations is set to $I_{max} = 1000$ and the set of initial probabilities for the strategy profile $\pi = \{\pi_1, \pi_2\}$ contains all values in $(0.1, 0.9)$ with a step size of 0.1. The learning parameter of the automaton is set to $\lambda = 0.01$, and the linear reward-inaction scheme is used for training.

For the three-strategy games, the maximum number of iterations is set to $I_{max} = 5000$ and the set of initial probabilities for the strategy profile $\pi = \{\pi_1, \pi_2\}$ contains only a reasonable small collection of values in $(0.1, 0.9)$ due to visibility. The learning parameter of the automata is set to $\lambda = 0.01$, and the linear reward-inaction scheme is again used for training.

For IA, the parameters are set to $\alpha = 0.6$ and $\beta = 0.3$, respectively [4]. IA is applied in two ways: *(1)* to both proposers, and *(2)* to one proposer and the responder. The first approach is used to compare the learning dynamics of FALA with and without IA in the two-strategy and three-strategy games. The second approach is additionally examined in the two-strategy game to contrast the impact of IA on the behavior of the proposers.

6.3 Results

In this section, we provide the results obtained by learning FALA with and without IA to play the two reduced versions of the CT game, presented in Section 5.

Figure 2 shows the learning dynamics of FALA with and without IA in the PD version of the two-player, two-strategy game. Without IA, the automata evolve to the mutual defection strategy *(L,L)*, while with IA, they evolve to the mutual cooperation strategy *(H,H)*. With regard to the time of convergence, the trajectories displayed in Figure 2 highlight that automata with IA converge very fast compared to those without IA.

The learning dynamics of FALA with and without IA in the SH version of the two-player, two-strategy game are displayed in Figure 3. Without IA, the automata evolve either to the mutual defection strategy or to the mutual cooperation strategy, depending on their initial probabilities. With IA, they evolve to the mutual cooperation strategy. With regard to the time of convergence, automata with IA again converge very fast compared to those without IA, as can be seen in Figure 3.

Figure 4 illustrates the dynamics of FALA without IA in the four versions of the two-player, three-strategy game for a small sample of the complete space of initial probabilities. The automata always convergence to one of the equilibria in their particular games.

Displayed in Figure 5 are the dynamics of FALA with IA in the two-player, three-strategy game. In all four games, the automata converge very fast to the mutual strategy *(H,H)* which is the one with the highest payoffs on the diagonal of the matrix.

Figure 6 illustrates the trajectories of the learning automata in the PD and the SH in the case of IA applied to one proposer and the responder in the two-strategy game. For the majority of initial conditions, the automata do not converge to an equilibrium point, but to the strategy which increases the payoff for the responder, i.e. *L*.

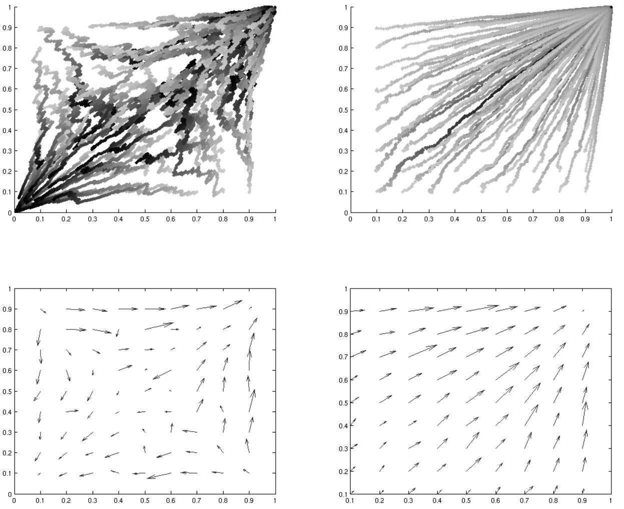
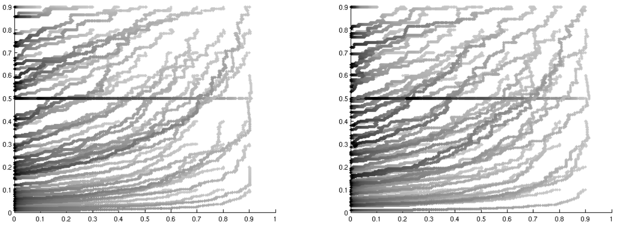


Figure 3: Trajectory plots of FALA in the Stag Hunt. The plots in the left column display trajectories of automata without inequity aversion, while the plots in the right column display those with inequity aversion.



(a) Prisoner's Dilemma.

(b) Stag Hunt.

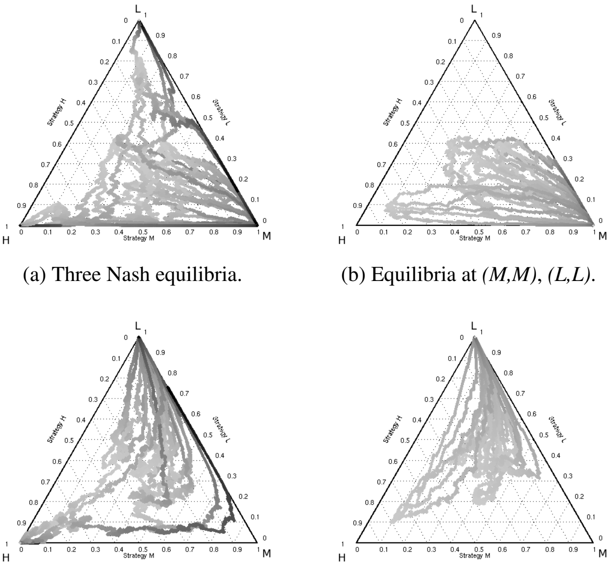
Figure 6: Trajectory plots of FALA with IA applied to one proposer and the responder in the two-strategy games.

6.4 Discussion

The results obtained from the experiments done in this study indicate convergence to higher payoffs and lower convergence time for automata with IA than for those without IA.

The reason behind the two effects mentioned above is that IA reduces the payoffs for game situations where one player gets a payoff of 0. The remaining game situations where both players get the same payoffs are not affected by IA; they stay the same.

When IA is applied, the general two-player, two-strategy game, given in Table 5 in Section 5, can be transformed to the payoff matrix shown in Table 10.



(a) Three Nash equilibria.

(b) Equilibria at *(M,M)*, *(L,L)*.

(c) Equilibria at *(H,H)*, *(L,L)*.

(d) Equilibrium at *(L,L)*.

Figure 4: Trajectory plots of FALA without IA in the three-strategy game.

	H	L
H	$\frac{1}{2}A, \frac{1}{2}A$	$0, B - \beta B$
L	$B - \beta B, 0$	$\frac{1}{2}B, \frac{1}{2}B$

Table 10: Transformed payoff matrix for the reduced two-player Colored Trails game with inequity aversion ($A > 1$).

Considering the PD, the highest payoff is B since it is assumed that $\frac{1}{2}A < B$. Therefore, this payoff is set to $B = 1$, if the payoffs are normalized. This normalization allows us to construct a general form of the PD that includes IA which is displayed in Table 11.

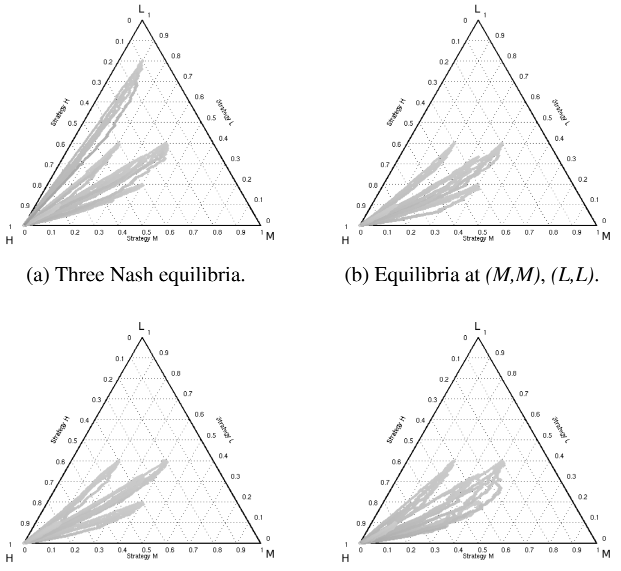
	H	L
H	$\frac{1}{2}A, \frac{1}{2}A$	$0, 1 - \beta$
L	$1 - \beta, 0$	$\frac{1}{2}, \frac{1}{2}$

Table 11: Transformed payoff matrix for the reduced two-player Colored Trails game with inequity aversion.

Table 11 illustrates that the payoffs in game situations where the players select different actions are only dependent on the parameter β. Since $A > B$ and thus $\frac{1}{2}A > \frac{1}{2}B$, the game transforms from a PD to a SH for $\beta \geq \frac{1}{2}$, with the typical two equilibria at mutual cooperation and mutual defection[4].

Contrasting the impact of the two approaches of IA, i.e. to both proposers or to one proposer and the responder, the proposers emerge to the mutual cooperation strategy in the former case and to the defection strategy in the latter case. In terms of IA, this result emphasizes that a proposer can either be fair to the other proposer or to the responder.

[4] This is a general result. In the experiments, only $\beta = 0.3$ is used.



(a) Three Nash equilibria.

(b) Equilibria at *(M,M)*, *(L,L)*.

(c) Equilibria at *(H,H)*, *(L,L)*.

(d) Equilibrium at *(L,L)*.

Figure 5: Trajectory plots of FALA with IA in the three-strategy game. The indicated equilibria are those of the *original* game; the equilibrium shifts to *(H,H)* for IA.

For the second approach of IA, i.e. to one proposer and the responder, the payoffs given to the responder in the two-strategy game are depicted in Table 12. It is assumed that the value of the payoffs are the same as for the proposers, thus the same constants are used.

	H	L
H	0	A
L	A	A

Table 12: Payoff matrix for the responder in the two-strategy Colored Trails game.

The changed payoffs for the first proposer (the first player) in the two-strategy game are given in Table 13.

	H	L
H	$\frac{1}{2}A - \beta\frac{1}{2}A$	0
L	$B - \alpha(A - B)$	$\frac{1}{2}B - \alpha(A - \frac{1}{2}B)$

Table 13: Payoff matrix for the proposer in the two-strategy Colored Trails game with inequity aversion for one proposer and the responder.

Since the two-strategy game shown in Table 5 is symmetric, the payoffs for the second proposer (the second player) are the same at the mutual cooperation and the mutual defection strategies, and the opposite at the other two strategies. This symmetry allows us to analyze the game with respect to the α- and β-parameters. Again, if payoffs are normalized, the reward B is set to $B = 1$. The changed payoffs for the first proposer are given in Table 14.

	H	L
H	$\frac{1}{2}A - \beta\frac{1}{2}A$	0
L	$1 - \alpha(A - 1)$	$\frac{1}{2} - \alpha(A - \frac{1}{2})$

Table 14: Transformed payoff matrix for the proposer in the two-strategy Colored Trails game with inequity aversion for one proposer and the responder.

The payoff for *(L,H)* can be rewritten as $1 - \alpha A - \alpha$ and the payoff *(L,L)* can be rewritten as $\frac{1}{2} - \alpha A - \frac{1}{2}\alpha$. Thus the payoff for *(H,L)* turns out to be higher than the payoff for *(L,L)*. Since the game is symmetric, this inequity also holds for the payoffs to the second proposer in the case of *(H,L)* and *(L,L)*. This is the reason why the learning automata do not converge to the equilibrium point at *(L,L)* for all initial conditions, and instead learn to play strategy L, when IA is applied to one proposer and the responder.

7 Conclusions

In this section, we discuss the research questions given in Section 1.1 and propose recommendations for further research.

7.1 Discussion of research questions

With respect to the main problem statement, i.e. *How can meaningful metastrategies for Colored Trails be established?*, the research we performed in this study underlines that, based on a number of sensible abstractions, we are able to derive a variety of metastrategies to play the game in an optimal way. From two extreme strategies for the proposer, i.e. to focus on its own gain or to focus on the responder's gain, an arbitrary number of intermediate strategies can be derived. The use of learning automata illustrates how agents could learn to optimally play Colored Trails using those metastrategies.

Considering the first research question, i.e. *Which abstractions or generalizations need to be made?*, we made abstractions with respect to the rational behavior of the players in the game, as described in Section 5. These abstractions scale the space of possible actions down to a set of metastrategies which can be applied to play the game in a reasonable way.

The second research question asked for the games onto which CT could be modeled. The research we performed in this study suggests that, using metastrategies, CT can indeed be mapped onto existing games. This conclusion is illustrated by the two-player games developed in Section 5. Starting with a simple model of the game with only two metastrategies and continuing with a more advanced model of a three-strategy game, we finally end up at a model for an arbitrary number of metastrategies.

As shown in Section 5, for a two-player, two-strategy version, the requirements of the SH or the PD can be satisfied depending on the values of the parameters used in the game. For a two-player, three-strategy version, four different three-strategy extensions of the original SH and the original PD evolve. In general, if CT is played with n metastrategies, it turned out to be representable as a two-player, n-strategy game with n potential Nash equilibria.

With respect to the third research question, i.e. *What is the role of inequity aversion in Colored Trails?*, we conclude that proposers can either be *fair* to each other or to the responder. By mutually cooperating with each other, they increase their own chance of winning the game. By defecting, they increase the responder's gain and decrease their own chance of winning.

7.2 Recommendations for further research

Given the performance of learning automata in the reduced versions of the CT game, research with learning automata in the original version of CT based on the results of this study should be performed.

Apart from CT, the effect of IA on the learning dynamics of other automata or learning schemes in the PD and the SH as well as in the three-player versions of those games, as developed in this study, could be investigated.

Besides, the effects of applying IA to the complete set of players, i.e. both proposers and the responder, could be examined.

References

[1] Axelrod, R. (1984). *The Evolution of Cooperation.* Basic Books.

[2] Chevaleyre, Y., Dunne, P., Endriss, U., Lang, J., Lemaître, M., Maudet, N., Padget, J., Phelps, S., Rodriguez-Aguilar, J., and Sousa, P. (2006). Issues in multiagent resource allocation. *Informatica*, Vol. 30, pp. 3–31.

[3] Jong, S. de (2009). *Fairness in Multi-Agent Systems.* Ph.D. thesis, Maastricht University.

[4] Fehr, E. and Schmidt, K.M. (1999). A theory of fairness, competition and cooperation. *The Quaterly Journal of Economics*, Vol. 114, No. 3, pp. 817–868.

[5] Gal, Y., Grosz, B.J., Kraus, S., Pfeffer, A., and Shieber, S. (2005). Colored trails: a formalism for investigating decision-making in strategic environments. *IJCAI Workshop on Reasoning, Representation, and Learning in Computer Games*, Edinburgh, Scotland.

[6] Gintis, H. (2001). *Game Teory Evolving: A Problem-Centered Introduction to Modeling Strategic Interaction.* Princeton University Press.

[7] Haim, G., Gal, Y., Kraus, S., and Blumberg, Y. (2010). Learning human negotiation behavior across cultures. *HuCom10 - Second International Working Conference on Human Factors and Computational Models in Negotiation*, Delft, The Netherlands.

[8] Hennes, D., Tuyls, K.P., Neerincx, M.A., and Rauterberg, G.W.M. (2009). Micro-scale social network analysis for ultra-long space flights. *The IJCAI-09 Workshop on Artificial Intelligence in Space*, Pasadena, California, USA.

[9] Hofstadter, D.R. (1983). Metamagical themas: Computer tournaments of the prisoner's dilemma suggest how cooperation evolves. *Scientific American*, Vol. 248, No. 5, pp. 16–26.

[10] Leyton-Brown, K. and Shoham, Y. (2008). *Essentials of Game Theory.* Morgan and Claypool.

[11] Narendra, K. and Thathachar, M. (1989). *Learning Automata: An Introduction.* Prentice-Hall International.

[12] Skyrms, B. (2004). *The stag hunt and the evolution of social structure.* Cambridge University Press.

[13] Wooldridge, M. (2009). *Multiagent Systems.* John Wiley and Sons Ltd.