False information and the emergence of conflict

Steven de Jong[ab]

[a]Computational Modeling Lab, Vrije Universiteit Brussel, Belgium
[b]Dep. of Knowledge Engineering, Maastricht University, The Netherlands

Abstract

Humans are not perfectly rational. In addition to considering rational arguments for performing certain actions, they also are influenced by, e.g., social considerations. In previous work, we computationally modelled the beneficial effects of this phenomenon, for use in multi-agent systems. Here, we explore disadvantageous effects of being only partially rational. More precisely, we aim to model how false information may cause humans to engage in an unnecessary conflict situation. We let participants play the Stag Hunt Game against virtual opponents, which are divided in two teams. After a number of games, false information (reputation) concerning one of the teams is given. We find that our participants are sensitive to this false information, even though it is verifiable that the information is indeed false. Integrating this phenomenon in existing or new computational models allows us to analyse why certain conflict situations occur, and to prevent them from occuring again.

1 Introduction

The goal of our previous research was to establish computational agreement and cooperation in social dilemmas (De Jong, 2009, De Jong and Tuyls, 2009, De Jong et al., 2008a,b). To this end, we looked at human behavior, focusing on human fairness. Humans are not completely individually rational: most notably, they are willing to take disciplinary action (e.g., altruistic punishment) if they consider others to act in an unfair manner (Fehr and Schmidt, 1999). Such disciplinary action leads to the establishment of agreement and cooperation in society, especially if we allow humans to spread reputation among their peers. We demonstrated that we may achieve a similar result in the context of multi-agent systems playing social dilemmas.

In our current research, we investigate the *disadvantageous* effects of humans being only partially individually rational, and being highly influenced by factors such as reputation. Our main aim is to predict the emergence of *conflict*. In accordance with existing work (Cederman and Girardin, 2007, Lim et al., 2007) on this matter, we wish to establish computational models, allowing us to predict where and when a conflict may emerge. In this paper, we present some initial steps in this research. We investigate the emergence of conflict between humans, as caused by two factors, i.e., (1) a heterogeneous society and (2) false information (more precisely, reputation). Using the results obtained in this paper, we will be able to extend existing computational models of, e.g., fairness, reputation and trust (Huynh et al., 2006), by incorporating the effect of these two factors, and matching models' predicted behavior with observed human behavior.[1] In subsequent research, we will investigate numerous other causes for conflict, e.g., resource scarcity.

Section 2 outlines how we model the emergence of conflict, as caused by the two factors, in a game-theoretic interaction named the Stag Hunt Game. Section 3 presents the main results we obtained in our experiments with humans, using the Stag Hunt Game. Section 4 provides pointers for modelling our results computationally, as well as presents a selection of related work. Section 5 concludes the paper.

[1]We note that this paper does not intend to (a) perform experiments with humans in a manner that is as well-controlled as experiments typically done by psychologists and behavioral economists, nor to (b) present a complete computational model of the effects of false information. The aim of the current paper is to show qualitatively that false information is a highly relevant factor contributing to the emergence of conflict. Future work should pursue to further examine human behavior and/or establish computational models.

2 Modelling conflict and false information

We model the (potential) conflict situation by means of the Stag Hunt Game (Skyrms, 2004), which is a repeated game. A possible payoff matrix for this game is given here.

	Stag	Hare
Stag	(**4**,4)	(**0**,2)
Hare	(**2**,0)	(**2**,2)

In every round of the game, a player needs to decide between playing the optimal action (Stag), which requires *trust* in the other player's willingness to play Stag as well, or playing a safer, but less rewarding action (Hare). Rationally, given the payoff matrix of the game as well as previous experience with a certain player (i.e., knowing which mixed strategy this player employs), the optimal response in the Stag Hunt Game is to play Stag whenever Stag is played by the other player in more than half of the games.[2]

Many researchers have used the Prisoners' Dilemma for investigations concerning human behavior in games where purely rational behavior may not be sufficient (Axelrod, 1984, Selten and Stoecker, 1986). However, this game lacks the coordination aspect present in the Stag Hunt; regardless of the strategy of the opponent, one should defect in the Prisoners' Dilemma to obtain the best result; the defective equilibrium is Pareto-dominated by the cooperative strategy, but the latter is not an equilibrium. This implies that the Prisoners' Dilemma basically requires participants to start a conflict. The Stag Hunt has two distinct equilibria instead of only one, where the cooperative equilibrium Pareto-dominates the defective one. Essentially, introducing altruistic punishment in interactions like the Prisoners' Dilemma (e.g., also the Public Goods Game; an approach we followed in previous work), may transform such interactions to Stag Hunt games.

In our experiments, we let humans play Stag Hunt Games against a set of *virtual* opponents. The probability that the opponents play Stag is set to 66%, which implies that players maximize their expected payoff by always playing Stag. We create two incentives for conflict (i.e., for prefering to play Hare).

1. The opponents are divided in two teams, which we call the Blue and the Red team. In this way, we create *heterogeneity*. Participants in the experiments are told they are a member of the Blue team, and that each team plays Stag with a certain fixed probability. Nothing is said about whether this probability is larger for Red or for Blue (in fact, the probabilities are identical, namely 66%). We use a fixed order in our games, i.e., every participant plays against the same sequence of opponents, with the same strategies.Participants play 10 practice games against each of the teams, which could allow them to find out that both teams are equally inclined to play Stag. In 10 subsequent 'real' games, we investigate whether the information about the teams leads to a bias in the participants' strategies against these teams. We call these 10 games the *pre-reputation* phase.

2. After the 20 practice games and 10 'real' games, we introduce *false information*, i.e., the participants are told that another member of the Blue team gives them the information that the Red team plays Stag with a probability of only 33%, instead of the actual probability of 66%; in behavioral literature, such information is often named reputation (Fehr, 2004). If this information were true, players should switch to playing Hare. Since participants have played 15 games against the Red team by the time this information is revealed, they can be able to detect that the information is false. We investigate whether the participants become less inclined to play Stag after the information is given to them. In addition, we establish whether this (erratic) behavior is corrected over time. To this end, participants play 20 more games after the false information is revealed; 10 games in the *post-reputation phase*, and 10 games in the *post-post-reputation phase*.

Participants are motivated to maximize their profit by means of a competition; the three players that accumulated the largest payoff in the course of the 30 'real' games, are awarded a small prize. The experiments are performed using an online survey. We manually removed the data of participants that clearly participated multiple times. Obviously, an online survey cannot be guaranteed to be free of cheating. We note that, if people indeed cheated, this will actually decrease the observable effects of the false information.

[2]Given that the other player plays Stag with a probability p and Hare with a probability $1-p$, playing Stag results in an expected payoff of $4p + 0(1-p)$, whereas Hare guarantees a payoff of 2. Thus, the expected payoff for playing Stag exceeds the expected payoff for playing Hare whenever $p > \frac{1}{2}$.

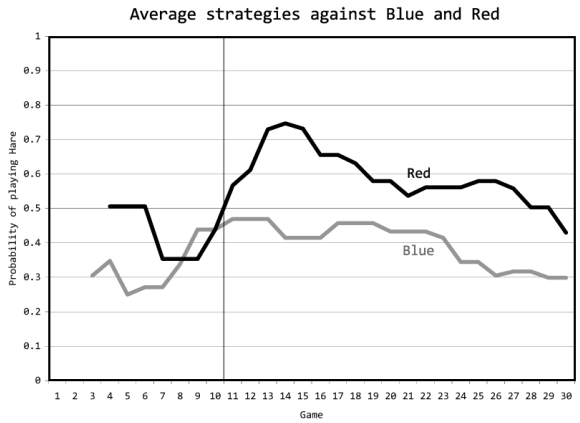

Figure 1: Average strategies of players against the two teams, over time. The vertical line denotes the time step in which the false reputation is revealed.

3 Experimental findings

Our survey was announced on various social network websites. After two weeks, 164 participants submitted results that were usable for analysis, i.e., their actions in the 30 games against our virtual opponents. We removed a number of results that were clearly obtained by participants that did not pay any attention, as well as results that were clearly obtained by participants that completed the survey more than once. We analysed our results, with the goal of determining whether the two incentives for conflict indeed caused conflict. We present two results here.

3.1 Average strategy over time

When we look at participants' strategy over time against the two teams (see Figure 1), we find two interesting observations. The figure shows a moving average of the participants' strategy, using a window of three games. In other words, the value at timestep t indicates the average strategy (of the 164 human participants) in the last three games before (and including) the game at timestep t. The first interesting observation concerns an initial bias with respect to the teams; during the pre-reputation phase, the Red team seems to be treated less cooperatively than the Blue team. Since the number of games that participants have played is rather small (i.e., five games per team), we cannot say with certainty that this bias is systematic. The second interesting observation is more prominent and also statistically significant: after the (false) information/reputation about the Red team has been revealed, participants become significantly less cooperative against the Red team. There is no significant change in participants' average attitude toward the Blue team. After an additional number of games, the attitude toward the Red team seems to be somewhat corrected. Details are given in Table 1; we provide the average percentage of defection, as well as the variance.

Against team	**Pre-reputation**	**Post-reputation**	**Post-post-reputation**
Blue	33% (22%)	43% (32%)	35% (30%)
Red	43% (34%)	67% (27%)	51% (32%)

Table 1: Average percentage of defection during the three phases of the experiment. We note that the (expected) optimal strategy in each phase, and against both teams, would be 0% defection.

3.2 A more detailed look at the strategy change

Examining participants' average strategy, we determined that the false information leads to a significant change in how the Red team is approached. On average, participants become significantly more defective toward the Red team. However, we also observe a high variance (see Table 1). It appears that, even though on average, participants tend to respond to the false information, the manner in which they respond varies

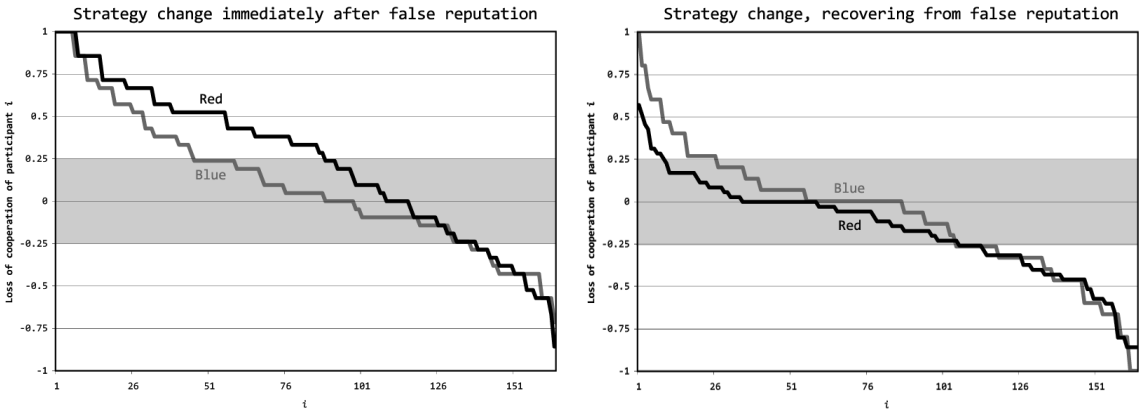

Figure 2: How participants change their strategies immediately after the false reputation (left), and how this change is later corrected.

between individual participants. An analysis of the *distribution* of participants' behavior is therefore appropriate. We took the following approach.

1. For every participant i, and for each of the two opponent colors, we calculated the average strategies μ_1^i for the pre-reputation phase (games 1–10), μ_2^i for the post-reputation phase (games 11–20), and μ_3^i for the post-post-reputation phase (games 21–30). For instance, $\mu_1^i = 0.6$ indicates that participant i played Hare 6 out of the 10 games in the pre-reputation phase.

2. For every participant, we calculated $\mu_{12}^i = \mu_2^i - \mu_1^i$ and $\mu_{23}^i = \mu_3^i - \mu_2^i$. If $\mu_{12}^i > 0$, this means that the participant became *more defective* in the post-reputation phase in comparison to the pre-reputation phase. Similarly, $\mu_{23}^i < 0$ means that this behavior (which is false) is corrected during the post-post-reputation phase, i.e., a while after the false information arrived.

3. We sorted the participants descendingly on μ_{12}^i and plot μ_{12}^i for each participant (see Figure 2, left). We did the same for μ_{23}^i (see Figure 2, right).

The two graphs in Figure 2 give us valuable information concerning how participants changed their strategies as a response to reputation. Obviously, since participants played only 10 games per phase of the experiment (i.e., pre-reputation, post-reputation, and post-post-reputation), only rather large changes in strategy can be considered to be meaningful. For this purpose, we decided not to consider strategy changes of 25% or less. We note that 25% is a rather arbitrary percentage; however, choosing any other percentage would not qualitatively change our observations. For clarity, the associated area of the graphs in Figure 2 is given a light gray background.

Looking at the leftmost graph in Figure 2, we see that 88 of the 164 participants become significantly more defective against Red opponents. In contrast, only 45 of the 164 participants have this behavior against Blue opponents. There are also participants that become more cooperative after the false information arrives. The difference between Red and Blue is not clearly visible here (27 participants became significantly more cooperative against Red, 25 against Blue). Thus, the clearest effect of the false information concerning the Red team is that a majority of participants take the false information into account in their strategy.

The rightmost graph in Figure 2 does not seem to show clear differences between strategies against Red and Blue opponents. However, there are differences that are worth considering. In case of the Blue team, there is a small but insignificant tendency to become less cooperative (54 participants do this, in comparison to 27 doing the opposite – note that we only consider strategy changes of more than 25%). In case of the Red team, there is in fact a significant tendency to become more cooperative (only 9 participants become more defective, 60 become more cooperative). Thus, in the post-post-reputation phase, the negative effects caused by the false information disappear slowly.

4 Applications for and pointers to computational models

The findings of the experiment described above may be used in, e.g., computational models of conflict, aimed at predicting when conflict arises in a certain group of people. Example applications do not only include war-like scenarios (see, e.g., Lim et al., 2007), but also the analysis of processes in society.

We give a current example for the applicability of our results in The Netherlands. A relatively recent development is the increasing prominence of a political party with strong anti-Muslim sentiments. Interestingly, the party leader has been accused of spreading false information (over-estimates) regarding the number of Muslims in Europe and the percentage of Muslims that are convicted for crimes (Winia, 2009). Looking at the results of our experiments, we may conclude that consistently reporting over-estimated Muslim crime rates is an effective strategy to win support for an anti-Muslim party, even though potential voters are perfectly able to determine that actual crime rates are lower.

Related work in the computational modelling of conflict finds many interesting causes for conflict, e.g., a critical size of clusters of ethnicity (Lim et al., 2007). In our own previous work, we were looking at computational models of agreement and fairness, by means of looking at the human example (De Jong, 2009). For the current work, especially the 'failed efforts' of this previous work are interesting. We indeed identified various human mechanisms in the literature that alledgedly could lead to cooperative solutions, but in fact caused the population to diverge when they were applied in a computational model. De Jong et al. (2008b) discuss successful and failed mechanisms in detail. Most prominently, we found that over-representing the human desire for fair solutions may lead to failure to achieve cooperative outcomes. For instance, imagine a group of individuals interacting in a network (cf. De Jong et al., 2008b, Santos et al., 2006). If we allow individuals i to remove the connection between them and a relatively defective neighbor j (replacing it by a random neighbor k of this neighbor j, to ensure that the network stays connected), this allows relative cooperators to effectively isolate relative defectors. In order to participate in interactions again, relative defectors need to change their strategy and become more cooperative. Now, imagine we add one of two other ideas, which sound very human.

First, we allow the relative cooperator i to connect to a *chosen* neighbor k of j, instead of a random one. This turns out to give an unfairly large advantage to the relative cooperator i (Uyttendaele, 2008); he may now be able select a neighbor k that is even more cooperative than him, and profit from this, essentially exploiting k. Thus, in effect, instead of deterring individuals from defecting, we stimulate them to defect.

Second, we keep the reputation of individuals and spread it throughout the population. Although in theory reputation may be beneficial, we also saw in the above that false reputation is able to seriously disturb the ability of a population to be cooperative. One of the problems of spreading reputation, even if we assume that the reputation is in fact correct, is that reputation describes individuals' *past* behavior. If, for instance, individuals refuse to play with those that have a bad reputation, this may create an irreversible schism, as those that behaved badly will never be allowed to prove that they changed their behavior. Thus, we end up with a group of cooperative individuals, and a group of defecting individuals, which cannot communicate with eachother and therefore will never resolve their differences (De Jong et al., 2008b).

As a starting point for computationally modelling conflict as a result of false information, we may use existing models of reputation and trust, as for instance FIRE (Huynh et al., 2006), which may be equipped with mechanisms concerning false information, probably leading to a decrease in trust within the MAS.

5 Conclusion

The fact that humans are only partially rational has many positive effects, especially because our limited rationality allows us to display social behavior (De Jong, 2009, Fehr and Schmidt, 1999). However, there are also clear drawbacks. In this paper, we investigate one of these drawbacks, i.e., the human inability to be unaffected by demonstrably false information. We analyse the behavior of 164 human players in the Stag Hunt Game, before and after false information is revealed to them concerning their opponents in the game (i.e., false reputation). We observe a clear tendency to take this false information into account. With the passing of time, the effects of the false information slowly disappear. Our observations may be incorporated in computational models of human interaction, aimed at predicting and analysing phenomena related to conflict. In addition to increasing our understanding of how such phenomena work, computational models may also allow us to prevent undesirable phenomena from occuring in the future.

Acknowledgement

The online survey reported in this paper was implemented by Koen Jacobs, in cooperation with and under supervision of the author.

References

R. Axelrod. *The Evolution of Cooperation*. New York: Basic Books, 1984.

L.-E. Cederman and L. Girardin. Toward realistic computational models of civil wars. In *Proceedings of the annual meeting of the American Political Science Association*, 2007.

S. de Jong. *Fairness in multi-agent systems*. PhD thesis, Maastricht University, NL, June 2009.

S. de Jong and K. Tuyls. Learning to cooperate in a continuous tragedy of the commons. In *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1185–1186, 2009.

S. de Jong, K. Tuyls, and K. Verbeeck. Fairness in multi-agent systems. *Knowledge Engineering Review*, 23(2):153–180, 2008a.

S. de Jong, S. Uyttendaele, and K. Tuyls. Learning to reach agreement in a continuous ultimatum game. *Journal of Artificial Intelligence Research*, 33:551–574, 2008b.

E. Fehr. Don't lose your reputation. *Nature*, 432:499–500, 2004.

E. Fehr and K. Schmidt. A Theory of Fairness, Competition and Cooperation. *Quart. J. of Economics*, 114: 817–868, 1999.

T. Huynh, N. R. Jennings, and N. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

M. Lim, R. Metzler, and Y. Bar-Yam. Global Pattern Formation and Ethnic/Cultural Violence. *Science*, 317 (5844):1540–1544, 2007.

F. C. Santos, J. M. Pacheco, and T. Lenaerts. Cooperation Prevails When Individuals Adjust Their Social Ties. *PLoS Comput. Biol.*, 2(10):1284–1291, 2006.

R. Selten and R. Stoecker. End behavior in sequences of finite Prisoner's Dilemma supergames : A learning theory approach. *Journal of Economic Behavior & Organization*, 7(1):47–70, March 1986. URL `http://ideas.repec.org/a/eee/jeborg/v7y1986i1p47-70.html`.

B. Skyrms. *The Stag Hunt and Evolution of Social Structure*. Cambridge: Cambridge University Press, 2004.

S. Uyttendaele. Fairness and agreement in complex networks. Master's thesis, MICC, Maastricht University, 2008.

S. Winia. Wilders overdrijft, Van der Laan praat onzin. *Elsevier*, 16-06, 2009.