5. The Implications of Case-Based Reasoning in Strategic Contexts[♣]

Though analogy is often misleading, it is the least misleading thing we have.
−SAMUEL BUTLER

5.1. Introduction

Case-Based Reasoning (CBR) is a form of reasoning by analogy within a particular domain (Aamodt and Plaza, 1994; Nicolov, 1997). In the context of problem solving, analogy can be defined as the process of reasoning from a solved problem which seems similar to the problem to be solved (Doran, 1997). Thus, CBR basically consists of "solving a problem by remembering a previous similar situation and by reusing information and knowledge of that situation" (Aamodt and Plaza, 1994). The rationale is that if a solution turned out to be satisfactory when applied to a certain problem it might work in a similar situation too.

Case-based reasoners do not employ abstract rules as the basis to make their decisions, but instead use similar experiences they have had in the past. Such experiences are stored in the form of cases. A case is "a contextualised piece of knowledge representing an experience that teaches a lesson fundamental to achieving the goals of the reasoner" (Kolodner, 1993, p. 13). Thus, when a case-based reasoner has to solve a problem, she is reminded of a similar situation that she encountered in the past, of what she did then, and of the outcome that resulted in the recalled situation. She then uses that 'similar past case' as a basis to solve the problem in the present. Case-based reasoning generally consists of four main tasks (Aamodt and Plaza, 1994):

[♣] Some parts of the material presented in this chapter have been published in Izquierdo L.R., Gotts, N.M. and Polhill, J.G. (2004) "Case-based reasoning, social dilemmas, and a new equilibrium concept", *Journal of Artificial Societies and Social Simulation*, 7(3), and in Izquierdo, L.R. and Gotts, N.M. (2005) "The implications of case-based reasoning in strategic contexts", *Lecture Notes in Economics and Mathematical Systems* 564, pp. 163-174.

1. *Retrieve* the most similar case or cases. Generally a case in CBR is rich in information and quite complex. Thus, performing similarity judgements is often an integral part of CBR. Admittedly, the representation of cases used in this chapter is particularly simple and, consequently, similarity judgements are straightforward; this is so because the primary objective of this research is to study the strategic implications of processes of reasoning based on one *single* distinctive past experience (in contrast with rule-based systems), and issues relating case representation are not so crucial for our purposes. The simple representation of cases used here may mean that certain researchers find the reasoning processes investigated in this chapter too unsophisticated to be called CBR; Aamodt and Plaza (1994) say: "a feature vector holding some values and a corresponding class is not what we would call a typical case description" (because it is too trivial). Thus, it is worth noting that the term CBR is used in this chapter –in a wider sense than Aamodt and Plaza's– to denote a process of reasoning based on one *single* distinctive past experience, selected for its similarity to the current situation.
2. *Reuse* the information and knowledge in the retrieved case to solve the current problem. The retrieved knowledge cannot always be directly applied, so some adaptation is sometimes required.
3. *Revise* the proposed solution. This involves the evaluation of the proposed solution.
4. *Retain* the relevant information for the future – i.e. learn.

Case-based reasoning is often used as a problem-solving technique in domains where the distinction between success and failure is either fairly easy to make or is made externally. However, in decision-making contexts in general, the distinction between what is satisfactory and what is not can be far from trivial, and thus, the question of whether a particular decision used in the past should be repeated, or a new decision should be explored is crucial. This dilemma naturally gives rise to Simon's (1957) notions of satisficing, as noted by Gilboa and Schmeidler (2001).

An alternative to CBR would be a rule-based system. One could induce the appropriate generalisations (rules) from the cases, and, in this view, CBR can be seen as a postponement of induction (Loui, 1999). However, when dealing with systems that are adaptive themselves (in the sense that they are constituted by adaptive agents), the 'rules' of the system vary as the system evolves and therefore agents must frequently revise their perceptions about the system. This could be done by constantly updating the set of induced rules or by using CBR. Agents who use CBR store the original cases without building rules that summarise them. In that way, cases can suggest solutions even to ill-defined problems, such as those arising in social dilemmas, for which there may not be an adequate set of general rules.

Origins and use of case-based reasoning

CBR arose out of cognitive science research in the late 1970s (Schank and Abelson, 1977; Schank, 1982). Schank and Abelson (1977) proposed that the general knowledge that we gain from experience is encoded in episodic memory as "scripts" that allow us to set up expectations and inferences. New episodes are processed by using dynamic memory structures which contain the episodes that are most closely related to the new episode; this process is called "reminding". Schank (1982) develops the idea that, far from being an irrelevant artefact of memory, reminding is at the root of how we understand and how we learn. Reminding occurs during the normal course of understanding, or processing some new information, as a natural consequence of the processing of that information. He argues that "we understand in terms of what we already understood".

There are several psychological studies that provide support for the importance of CBR as problem-solving process in human reasoning, especially for novel or difficult tasks (see Ross (1989) for a summary). Klein and Calderwood (1988) studied over 400 decisions made by experienced decision makers performing a variety of tasks in operational environments and concluded that "processes involved in retrieving and comparing prior cases are far more important in naturalistic decision making than are the application of abstract principles, rules, or conscious deliberation between alternatives". Drawing on their empirical

studies, they also developed a descriptive model of decision making in which the attempt is to satisfice rather than optimise.

More recently, Gayer et al. (2007) have empirically examined the *relative* importance of rule-based versus case-based reasoning in housing asking prices. They hypothesise on theoretical grounds that case-based reasoning has relatively more explanatory power in the rental apartment market, whilst rule-based reasoning is relatively more prevalent in the sales market, and they find empirical support for this hypothesis when tested with two databases (rentals and sales) of asking prices on apartments in the greater Tel-Aviv area. However, their interpretation of case-based reasoning is significantly different from that explained above. In their model, case-based reasoning is modelled using a similarity-weighted average that makes use of *all* cases available at the time of making a decision. In general terms, they conjecture that, in comparison to rule-based reasoning, case-based reasoning will be more prevalent in non-speculative markets than in speculative ones. They also state their belief that both modes of reasoning are likely to play a role in almost any decision-making process, and that a variety of factors may affect their relative importance.

It seems therefore that CBR is plausible as at least a partial representation of how people make use of past experience: that they recall circumstances similar to those they now face and remember what they did and with what outcome (see for example Kahneman et al., 1982).

There are also a number of industrial applications of CBR (Watson, 1997), particularly in domains where there is a need to solve ill-defined problems in complex situations; in such situations, it is difficult or impossible to completely specify all the rules (if they exist at all) but there are cases available.

Within the domain of theoretical economics, a Case-Based Decision Theory (CBDT) has been proposed by Gilboa and Schmeidler (1995; 2001). CBDT is a formal theory of decision based on past experiences which was initially inspired by case-based reasoning. Having said that, as noted by the authors, CBDT has not much in common with CBR beyond Hume's basic argument that "from causes

which appear similar we expect similar effects". As pointed out when describing the empirical study conducted by Gayer et al. (2007), the main difference between CBR and CBDT is that while a defining feature of CBR is that "thought and action in a given situation are guided by a single distinctive prior case" (Loui 1999), in CBDT decision-makers rank available acts according to the similarity-weighted sum of utilities that resulted in *all* available cases. For the formalisation of an assessment rule based on such a similarity-weighted function see Gilboa et al. (2006). In any case, Gilboa and Schmeidler (1995; 2001) do not see case-based decision theory (CBDT) as a substitute for expected utility theory (EUT), but as a complement. They argue that CBDT may be more plausible than EUT when dealing with novel decision problems, or in situations where probabilities cannot easily be assigned to different states of the world (uncertainty, as opposed to risk), or if such states of the world cannot be easily constructed (ignorance). They also highlight that CBDT naturally gives rise to the notions of satisficing decisions and aspiration levels.

Pazgal (1997) and Kim (1999) apply CBDT in strategic contexts. Pazgal (1997) analyses general games of mutual interest (i.e. games where there exists a unique pure strategy profile that gives the highest possible payoff to every player), and Kim (1999) focuses on symmetric 2x2 games of mutual interest to study the aspiration updating mechanism in greater depth[18]. The decision-making algorithm employed by players in these two studies bears very little resemblance to CBR as interpreted above: players in Pazgal's and Kim's models do not consider different cases or experiences, they choose the action that has given them the highest cumulative past payoff (relative to their current aspiration) throughout the whole history of the game, and their aspiration thresholds are updated using a weighted average of its previous value and an average function of received payoffs. This

[18] Kim (1999) studies 2x2 games with an outcome (i.e. a *pure strategy profile*) which every player strictly prefers, and refers to these as "common interest" games. Following Aumann and Sorin (1989), I use the term "common interest game" to denote the wider class of games where there is a unique *payoff profile* that strongly Pareto dominates all other payoff profiles (and this payoff profile may be achieved via several strategy profiles), and I use the more specific term "mutual interest game" to denote games where there exists a unique *pure strategy profile* that gives the highest possible payoff to every player.

decision-making algorithm (identified by the authors as a form of case-based maximisation) is significantly different from that consisting in maximising average payoffs (as nicely illustrated by Kim (1999)), but it is also fundamentally different from CBR as interpreted in this chapter. As a matter of fact, it seems to us that the essence of these two models is closer to reinforcement learning than to case-based reasoning, as also noted by Bendor et al. (2001a; 2001b).

To our knowledge, the implications of CBR interpreted as explained above in strategic contexts had never been formally explored up until now. In this chapter we develop and analyse a game theoretical model where individuals use a very simple form of CBR.

Structure of this chapter

In this chapter we use social dilemma games to illustrate the strategic implications of case-based reasoning. The following section is devoted to explaining why social dilemmas in particular are especially revealing to understand the differences between reasoning by cases and reasoning by rules. Section 5.3 presents a simple model that is used to shed light on the conditions under which CBR as individual decision mechanism may entail cooperation in social dilemmas. The results obtained with this model are presented and discussed in sections 5.4 and 5.5 respectively. Section 5.6 presents a generalisation of the model analysed in sections 5.4 and 5.5. In particular, players in the more general model may make occasional mistakes in their decisions. The dynamics of this second model are explained and discussed in 5.7. Finally, section 5.8 presents the conclusions of this chapter.

5.2. Case-based reasoning and social dilemmas

This chapter provides various results on the *asymptotic* dynamics of a rather general form of CBR for any finite normal-form game (see section 5.7). The *transient* dynamics of CBR models, however, strongly depend on the definition of the particular CBR algorithm employed by players and also on the specific game they play. Thus, to explore the whole dynamics of games played by agents who use a simple form of CBR, the scope of study has had to be limited to some

extent. In particular, whenever it has been found that the specific parameterisation of the game has made a difference I have focused on analysing social dilemmas.

Social dilemmas offer a promising arena to distinguish the differences between reasoning by cases (or outcomes[19]) and reasoning by rules (or strategies). The following illustrates why this is the case using the Prisoner's Dilemma. Although defining rational strategies in interdependent decision-making problems is by no means trivial, it seems sensible to assume that a) rational players choose dominant strategies[20], and b) rational players do not choose dominated strategies[21]. Similarly, even though defining rational outcomes cannot be done without controversy, it also seems sensible to agree that rational outcomes must be Pareto optimal[22]. Assuming only those necessary conditions for the rationality of strategies and outcomes, we can state that in the one-shot Prisoner's Dilemma and other social dilemmas, even though there is a clear causal link between strategies and outcomes, rational strategies (understood as those chosen by rational players) lead to outcomes that are not rational, whereas rational outcomes are generated by strategies that are not rational (i.e. those strategies that a rational player would never select).

In this chapter we explore two social dilemma games: a 2-player and an *n*-player version of the Prisoner's Dilemma (PD). Because of the players' decision making algorithms (explained in sections 5.3 and 5.7), the actual values of the payoffs are not relevant as long as they satisfy:

$$Temptation > Reward > Punishment > Sucker$$

---

[19] An outcome is a particular combination of decisions, each of them made by one player.

[20] Recall that, for a player A, strategy S_A is (strictly) dominant if for each combination of the other players' strategies, A's payoff from playing S_A is (strictly) more than A's payoff from playing any other strategy (Gibbons, 1992, p. 5).

[21] For a player A, strategy S_A is (strictly) dominated by strategy S^*_A if for each combination of the other players' strategies, A's payoff from playing S_A is (strictly) less than A's payoff from playing S^*_A (Gibbons, 1992, p. 5).

[22] An outcome is Pareto optimal if there is no other outcome where at least one player is better off and no player is worse off.

In the *n*-player social dilemma every player gets a reward as long as there are no more than *M* defectors ($M < n$). The payoff that defectors get is always higher than the payoff obtained by those who cooperate (*Def-P* > *Coop-P*). However, every player is better off if they all cooperate than if they all defect (*Coop-P* + *Reward-P* > *Def-P*). Figure 5-1 shows the payoff matrix for a particular player:

	Fewer than *M* others defect	***M* others defect**	**More than *M* others defect**
Player cooperates	***Coop-P* + *Reward-P***	***Coop-P* + *Reward-P***	***Coop-P***
Player defects	***Def-P* + *Reward-P***	***Def-P***	***Def-P***

Figure 5-1. Payoff matrix of the "Tragedy of the Commons game" for a particular agent.

This game has been called in the literature the "Tragedy of the Commons game" (Kuhn, 2001) after the influential paper written by Hardin (1968). Henceforth we will refer to this game as the TC game. When the maximum number of defectors *M* for which the reward is given is high, it represents a version of the "volunteer's dilemma" (Brenan and Lomasky, 1984; Diekmann, 1985): a group needs a few volunteers, but each member is better off if others volunteer. If the number of players is large enough, the case when exactly *M* others defect is sufficiently unlikely that for all intents and purposes it can be ignored. Assuming the latter, we have a "social dilemma" as defined by Dawes (1980): "all players have dominating strategies that result in a deficient equilibrium"[23]. In any case, we have a "problematic social situation" (Diekmann, 1986; Raub and Voss, 1986), or social dilemma in a broader sense, which can be defined in game theory terms as a game with Pareto inefficient[24] Nash equilibria. The TC game differs from the two-player PD in three important ways:

1. In the TC game, for a small number of players, the state of "minimally effective cooperation" (exactly *M* defectors) is not negligible, so there is not a dominant strategy.

[23] An equilibrium is deficient if there exists another outcome which is preferred by every player.

[24] An outcome is Pareto inefficient if there is an alternative in which at least one player is better off and no player is worse off.

2. In the TC game, using pure strategies, there are two Nash equilibria: everyone defecting (universal defection[25]) and exactly *M* defectors (minimally effective cooperation).
3. In the two-player PD, universal cooperation is a Pareto optimal outcome since no player can be better off without making the other player worse off. However, in the TC game the only Pareto optimal outcome is the state of minimally effective cooperation.

5.3. The CBR model

In this section we present a simple CBR decision-making algorithm that players will use to decide whether to cooperate or not when confronted with one of the two social dilemma games described in the previous section. This model will be named "the CBR model". Individuals play repeatedly the game – once per time-step – and every time they do so, each player retains a case (representing the experience they lived in time-step *t*) which comprises:

1. The time-step *t* when the case occurred.
2. The *perceived* state of the world at the beginning of time-step *t*, characterised by the value of the following descriptors in the preceding *ml* (for *memory length*) time-steps:
    - Descriptor 1 (D1): the number of other defectors.
    - Descriptor 2 (D2): the decision that the player holding the case made.

    As an example, if $ml = 2$ then the perceived state of the world for the case-holder will be determined by the number of other defectors and the decision she made, both in time-step $t - 1$ and in time-step $t - 2$).
3. The decision the case-holder made in that situation, *i.e*. whether she cooperated or defected in time-step *t*, having observed the state of the world in that same time-step.
4. The payoff that the case-holder obtained after having decided in time-step *t*.

Thus the case representing the experience lived by player *A* in time-step *t* has the following structure:

[25] Universal defection is a Nash equilibrium as long as $M < n-1$.

t	$df_{t-ml} \ldots df_{t-2} \quad df_{t-1}$ $d_{t-ml} \ldots d_{t-2} \quad d_{t-1}$	d_t	p_t

where

df_t is the number of defectors (excluding player *A*) in time-step *t*,

d_t is the decision made by player *A* in time-step *t*, and

p_t is the payoff obtained by player *A* in time-step *t*.

The number of cases that players can keep in memory is unlimited. It is also worth noting that no cases are available for any player until (ml + 1) time-steps have gone by in the simulation. Players make their decision whether to cooperate or not by retrieving two cases: the most recent case which occurred in a *similar* situation for each of the two possible decisions (*i.e.* each of the two possible values of d_t). A case is perceived by the player to have occurred in a *similar* situation if and only if its state of the world is a perfect match with the current state of the world observed by the player holding the case. The only function of the perceived state of the world is to determine whether two situations look *similar* to the player or not. In a particular situation (*i.e.* for a given perceived state of the world) a player must face one of the following three possibilities:

1. The player cannot recall any previous situations that match the current perceived state of the world. In CBR terms, the Agent does not hold any appropriate cases for the current perceived state of the world. In this situation the player will decide at random.
2. The player does not remember a previous similar situation when she made a certain decision, but she does recall at least one similar situation when she made the other decision. In CBR terms, all the appropriate cases the player recalls have the same value for d_t. In this situation, the player will explore the non-applied decision if the payoff she obtained in the last previous similar situation was below her *Aspiration Threshold AT*; otherwise she will keep the same decision she previously applied in similar situations.
3. The player remembers at least one previous similar situation when she made each of the two possible decisions. In this situation, the player will focus on the most recent case for each of the two decisions and choose the decision

that provided her with the higher payoff[26]. In this way, players adapt their behaviour according to the most recent feedback they got in a similar situation.

This completes the specifications of "the CBR model". The UML activity diagram of the players' decision making algorithm is outlined in Figure 5-2. In the simulation experiments reported in this chapter, all the players share the same aspiration threshold *AT* and the same memory length *ml*. These are the two crucial parameters in the CBR model, determining when an outcome is satisfactory and when two situations are similar, respectively. The behaviour of a slightly more advanced socioeconomic Agent which also uses CBR in their decision-making algorithm but takes into account social approval is explored in Izquierdo *et al*. (2003).

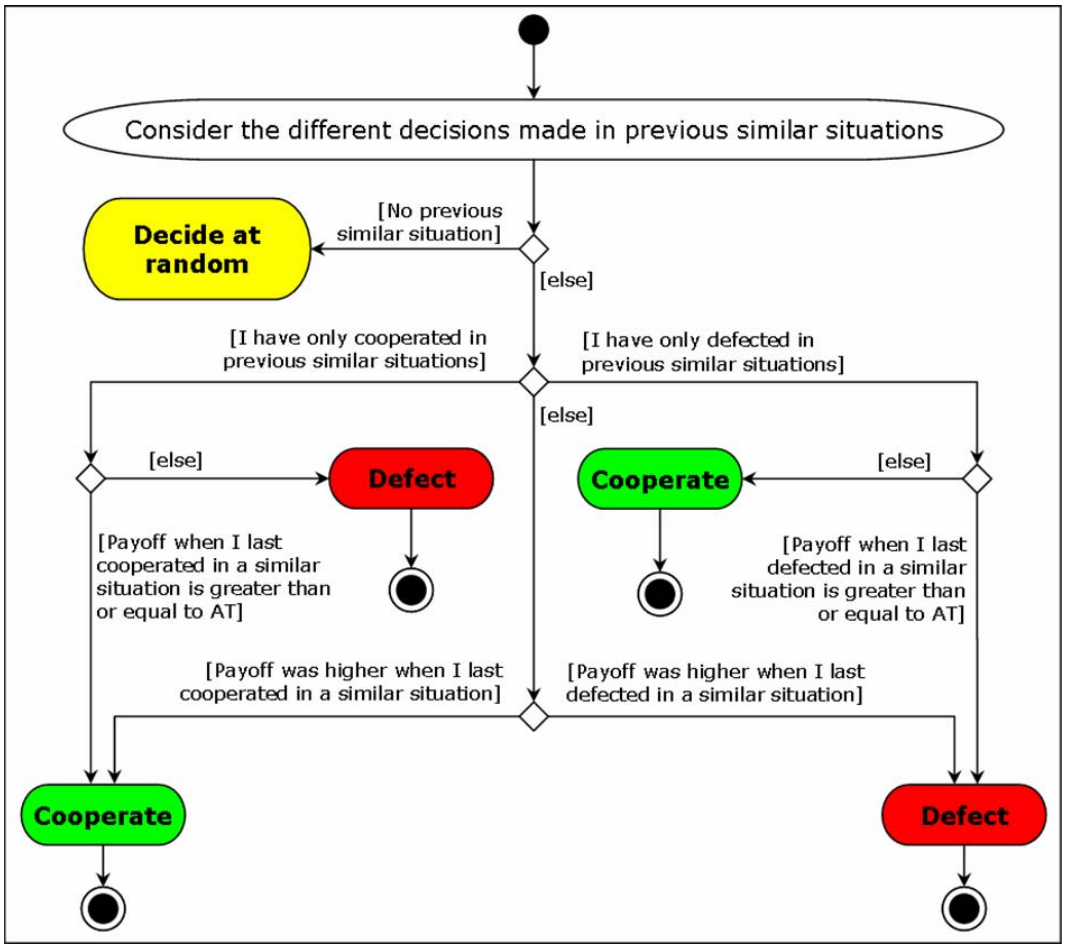

Figure 5-2. UML activity diagram of the CBR decision making algorithm.

[26] A tie is impossible in either of the two games analysed in this chapter.

5.4. Results with the CBR model

The software used to conduct the experiments reported in this section was written in Objective-C using the Swarm libraries (http://www.swarm.org) and is available in the Supporting Material together with a user guide under the GNU General Public Licence. The program is known to work on a PC using Swarm 2.1.1 and on a Sun Sparc using Swarm 2001-12-18.

As might be expected, the CBR model is very sensitive to the decisions that players make at random. Since the model has stochastic components, the results for a given set of parameters cannot be given in terms of assured outcomes but only as a range of possible outcomes, each with a certain probability of happening. The probability of each outcome can be either estimated by running the model several times with different random seeds or, under certain circumstances, exactly computed.

Players in the CBR model make decisions at random only when they perceive a novel state of the world. Since the number of different states of the world that a player can perceive is finite, so is the number of random decisions the player can make. Therefore, simulations must end up in a cycle. To study how often players cooperate in the PD we define the 'cooperation rate' as the number of times bilateral cooperation is observed in a cycle divided by the length of the cycle. Similarly, we define the 'reward rate' in the TC game as the number of times the reward is given in a cycle divided by the length of the cycle.

5.4.1. Prisoner's Dilemma

Aspiration Thresholds

It is important to realise that when players play the PD, they share the same perception of the state of the world (defined by the last *ml* moves of the two Players) in the sense that any two situations that look the same to one player will also look the same to the other player and any two situations that look different to one player will also look different to the other player. Therefore, at any given time in the simulation our players will have visited any given state of the world the same number of times. This shared perception of the state of the world means that, for a certain state of the world, the only relevant factor is the random decision that they make when they first experience that situation.

The decision dynamics for a certain state of the world are summarised in Table 5-1. Consider for now the first four rows of the table ($T < AT$). These represent the case where the aspiration threshold AT (for both players) exceeds T. The first time any particular state of the world occurs, both players will choose C (Cooperate) or D (Defect) at random (column headed "1st visit"). When the same perceived state occurs a second time, the responses will be as shown in the "2nd visit" column, and so on. The table shows that by the third visit to that state, either both players are cooperating or both players are defecting, and both will then continue to make the same response. The other four sets of rows in the table show what happens when the AT is in each of four lower ranges of values.

Aspiration Thresholds (AT)	1st visit (random)	2nd visit	3rd visit	4th visit and onwards	*x*	*y*
$T < AT$	CC	DD	CC	CC	1	-
	CD	DC	DD	DD	-	2
	DC	CD	DD	DD	-	2
	DD	CC	CC	CC	1	-
$R < AT \leq T$	CC	**DD**	CC	CC	1	-
	CD	**DD**	DC	DD	-	2
	DC	**DD**	CD	DD	-	2
	DD	**CC**	CC	CC	1	-
$P < AT \leq R$	CC	CC	**CC**	CC	0	-
	CD	DD	**DC**	DD	-	2
	DC	DD	**CD**	DD	-	2
	DD	CC	**CC**	CC	1	-
$S < AT \leq P$	CC	**CC**	CC	CC	0	-
	CD	**DD**	DD	DD	-	1
	DC	**DD**	DD	DD	-	1
	DD	**DD**	DD	DD	-	0
$AT \leq S$	CC	CC	CC	CC	0	-
	CD	CD	CD	CD	-	-
	DC	DC	DC	DC	-	-
	DD	DD	DD	DD	-	0

Table 5-1. Decisions made by each of the two players in the PD when visiting a certain state of the world for the *i*-th time. In the first column, payoffs are denoted by their initial letter. In columns 2 to 5, the first letter in each pair corresponds to the decisions of one player, the second letter to those of the other. C is cooperation and D is defection. The first imbalance between CC and DD for every value of AT has been shadowed. The meaning of x and y is explained in the text. The results shown in this table are independent of the memory length.

There are two states of the world that appear to be particularly important in the dynamics of the game. One is that where there have been *ml* successive bilateral cooperations (let us call it *mlBC*); the other is where there have been *ml* successive bilateral defections (let us call it *mlBD*). Whenever bilateral cooperation follows a visit to *mlBC*, then *mlBC* is immediately revisited (since players observe again that they both cooperated in the last *ml* time-steps). Similarly, whenever bilateral defection follows a visit to *mlBD*, then *mlBD* is immediately revisited (since players observe again that they both defected in the last *ml* time-steps). We can then define *x* as the number of times that *mlBC* has to be revisited after it has been abandoned before stable cooperation is reached, and *y* as the number of times that *mlBD* has to be revisited after it has been abandoned before stable defection is reached. As an example, when $AT > T$, if both players happen to cooperate when they observe *mlBC* for the first time, then they will both experience *mlBC* for the second time in the following time-step. Both of them will then defect (2^{nd} visit to *mlBC*), and in doing so will abandon *mlBC*. If *mlBC* is then revisited (3^{rd} visit), it will never be left again. In this hypothetical example, the number of times *x* that *mlBC* had to be revisited after it was abandoned before stable cooperation was reached was 1. This information is included in Table 5-1 and its significance will be explained later.

When the simulation locks in to a cycle (and it necessarily does), the states that make up the cycle are repeatedly visited, leading to outcomes shown in the "4^{th} visit and onwards" column in Table 5-1. Looking at that column, we can identify two values for the aspiration threshold *AT* that make a particularly important difference: *Sucker* and *Punishment*.

- When $AT > Sucker$, simulations lock in to cycles which are necessarily made up of bilateral decisions (both players cooperate or defect at the same time), since if a player receives the S*ucker* payoff in any situation, they will never cooperate again in that situation. In this sense our players are particularly unforgiving. Players with aspiration thresholds greater than *Sucker* cannot be systemically exploited. The importance of this will be discussed later.
- When $AT > Punishment$, there is a qualitative jump in terms of average cooperation rates. This is because if $AT > Punishment$, when both Players

defect the first time they experience a certain state of the world, they will end up cooperating in that state, but they will end up defecting if $AT \leq$ *Punishment*.

Taking into account the two previous points and looking at the "4th visit and onwards" column in Table 5-1, one could then think that average cooperation rates should be 25% if $AT \leq$ *Punishment* and 50% if $AT >$ *Punishment* regardless of the Memory Length, but one would be wrong. Figure 5-3 shows the importance of aspiration thresholds and how they can modify the effect of the memory length.

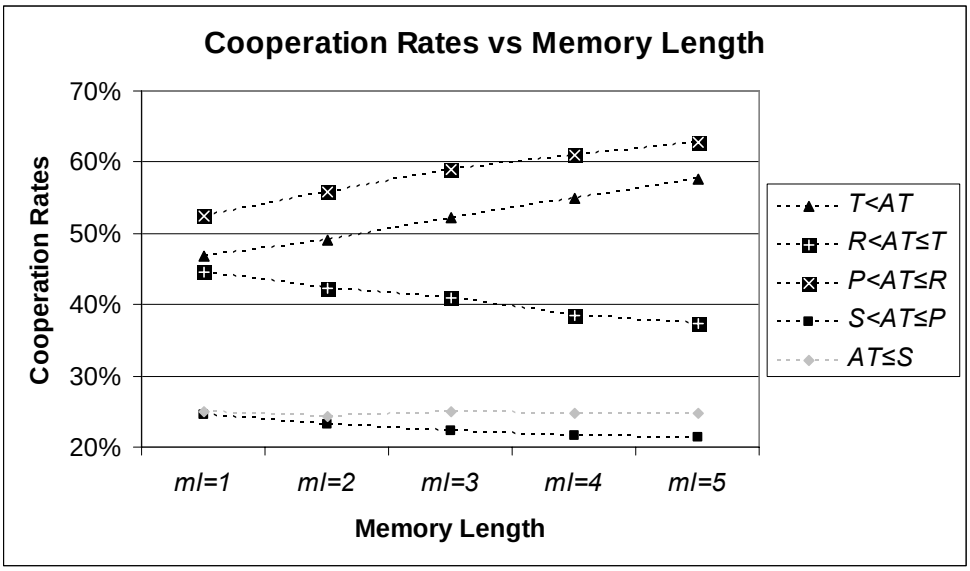

Figure 5-3. Average cooperation rates when modelling two players with Memory Length *ml* and Aspiration Threshold *AT*, playing the PD. The average cooperation rate shows the probability of finding both Players cooperating once they have finished the learning period (*i.e.* when the run locks in to a cycle). The values represented for *ml* = 1 have been computed exactly. The rest of the values have been estimated by running the model 10,000 times with different random seeds. All standard errors are less than 0.5 %.

The interactions between the aspiration threshold and the memory length can be explained by taking into account two factors. Both factors are related to the fact that, as the memory length increases, the number of possible perceived states of the world grows exponentially and it becomes less likely for any given state of the world to be revisited. From now on let us refer to each payoff by its initial letter.

1. The first factor concerns only the relative frequency of stable bilateral cooperation and stable universal defection[27]. This factor is present for any $AT > S$ and represents a bias towards cooperation. Looking at Table 5-1, one could expect stable bilateral defection to be three times more likely than stable bilateral cooperation if $S < AT \leq P$, and as likely as stable bilateral cooperation if $AT > P$. However, as the memory length increases, there is a certain bias towards stable bilateral cooperation. For the simulation to lock in to stable bilateral cooperation, it is required that a bilateral decision (a bilateral cooperation if $S < AT \leq P$) follows the first visit to the state of the world formed by *ml* bilateral cooperations (*mlBC*) and that the same state of the world *mlBC* is revisited *x* more times after it is abandoned; similarly, stable bilateral defection requires a unilateral decision (or bilateral defection if $S < AT \leq P$) following the first visit to the state of the world formed by *ml* bilateral defections (*mlBD*) and *y* more visits to that state of the world *mlBD* after it is abandoned. As we can see in Table 5-1, except for the trivial case[28] where $AT \leq S$, the average *x* is always less than the average *y* for any given aspiration threshold. For high values of the memory length, revisiting a state can take a very long time and the fact that stable bilateral cooperation needs fewer visits (*x*) to settle down than stable bilateral defection does (*y*) is an important bias towards the frequency of stable bilateral cooperation.

2. The second factor explains why average cooperation rates not only fail to increase, but actually decrease with memory length for $S < AT \leq P$ and $R < AT \leq T$. This factor is present for $S < AT \leq T$ and it represents a bias towards cooperation if $P < AT \leq R$, and a bias towards defection if $S < AT \leq P$ or $R < AT \leq T$. For any $AT > S$, the simulation ends up in a cycle of bilateral decisions. Therefore, it is crucial to study whether there is a bias towards cooperative bilateral decisions (CC) or towards defective bilateral decisions (DD) in the players' learning process. Table 5-1 shows the history of decisions made by the players having observed any particular state of the world for different aspiration thresholds. The first imbalance between CC and DD for every value of *AT* has

[27] This effect is explained in detail by Izquierdo et al. (2003).

[28] If the Aspiration Threshold does not exceed *Sucker*, Agents repeat the same decision that they made at random the first time they visited a certain state of the world whenever they visit the same state again.

been shadowed (*e.g.* if $S < AT \leq P$ the first imbalance occurs in the second visit, where DD is three times more likely to happen than CC). Imbalances in the earlier visits to a state of the world are more important because those in later stages might never materialise if a cycle is reached before they can occur. Imbalances in the component parts of the state of the world (CC and DD) make certain states of the world more likely to occur than others, hence leading to biases in the cooperation rate. What is not obvious is why the importance of such imbalances (in terms of reward rates) increases with the value of memory length. This is so because, even ignoring the fact that some states of the world are more likely to occur than others, not all states of the world are equally likely to form part of a cycle; some states can form cycles more easily than others[29], and their relative frequency depends on the memory length. This is certainly the case for *mlBC* and *mlBD*. Not only are they the only states of the world that can form cycles just by themselves (assuming $AT > S$), but they also need fewer revisits to settle than the rest of the possible states of the world (see previous paragraph). Roughly half of the simulation runs reported in this paper with $AT > S$ ended up in cycles made up by either *mlBC* or *mlBD*. This means that an imbalance between the frequency of *mlBC* and *mlBD* can affect the reward rate substantially. The imbalance between *mlBC* and *mlBD* given an imbalance between CC and DD does depend on the memory length. To clarify this, assume that DD is always z times more likely than CC; then *mlBD* will be z^{ml} times more likely than *mlBC*. This analysis is not a proof since successive states of the world are not independent, but it clarifies why imbalances gain importance as the value of the memory length increases. As we can see in Table 5-1, if $S < AT \leq P$ or $R < AT \leq T$, the imbalance is towards the defective bilateral decision, making *mlBD* more likely to occur relative to *mlBC* as memory length increases, and thus reducing the average cooperation rate. On the other hand, if $P < AT \leq R$, the imbalance is towards cooperation.

The summary of the effect of each of the two factors depending on the AT outlined above is shown in Table 5-2, together with the total effect found in the simulations. We have not yet proved that the two effects explained here are the only operating factors.

[29] Or, conversely, some cycles comprise fewer different states of the world than others.

	$AT \leq S$	$S < AT \leq P$	$P < AT \leq R$	$R < AT \leq T$	$T < AT$
Effect of factor 1	No bias	Bias towards cooperation	Bias towards cooperation	Bias towards cooperation	Bias towards cooperation
Effect of factor 2	No bias	Bias towards defection	Bias towards cooperation	Bias towards defection	No bias
…	…	…	…	…	…
Total effect	No bias	Bias towards defection	Bias towards cooperation	Bias towards defection	Bias towards cooperation

Table 5-2. Effect on average cooperation rates of each of the two factors outlined in the text above depending on the value of AT, and results from the simulation runs.

It is clear that in CBR, not only *what* is learnt, but the actual *process* of learning can be of major importance, and aspiration thresholds play a crucial role in that process. Consider, for example, the difference between the cases where $P < AT \leq R$ and where $R < AT \leq T$. In both cases, players will learn to cooperate in any given state of the world if they happen to make the same decision the first time they visit that state, and they will end up defecting in that situation otherwise. Therefore, for those two values of AT, we could expect average cooperation rates to be the same or at least similar. However, because the actual process of learning is different, differences in average cooperation rates are substantial and get larger as the memory length increases (see Figure 5-3).

Importance of a common perception of the state of the world

To study the importance of having a shared perception of the state of the world in the PD, we studied the outcome of the game when played by players with partial representations of the state of world: players who only look at the other player's actions (only descriptor D1) and players who only look at their own actions (only descriptor D2). In both these cases, the two players may perceive the state of the world differently. Figure 5-4 shows the results obtained for $AT > T$. The results for other aspiration thresholds are very similar[30] so they are omitted.

[30] Except, again, for the trivial case where $AT \leq S$, in which the average cooperation rate is always 25%.

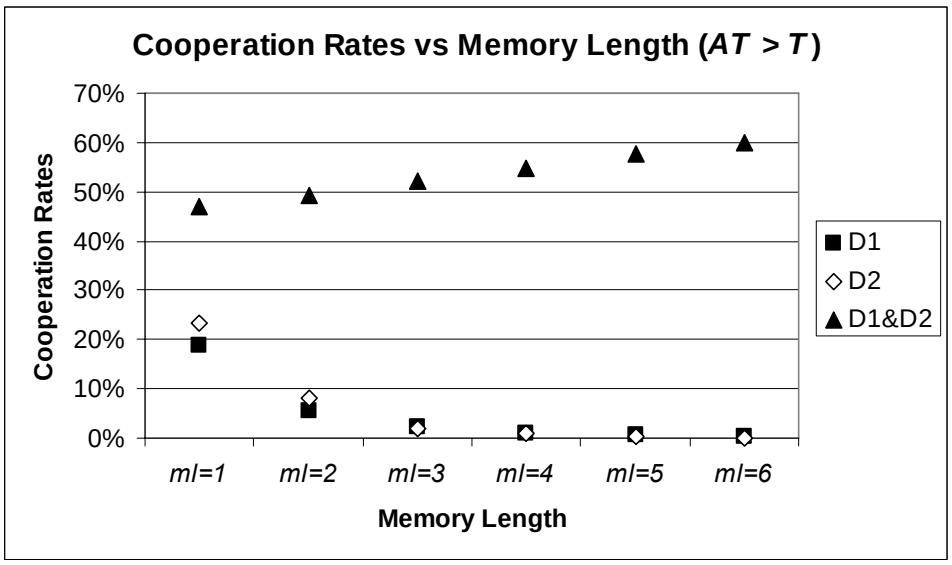

Figure 5-4. Average cooperation rates when modelling two players with Memory Length *ml*, Aspiration Threshold greater than *Temptation*, and with 3 different representations of the state of the world (D1, D2, and D1&D2), playing the PD. The values represented for *ml* = 1 have been computed exactly. The rest of the values have been estimated by running the model 10,000 times (*ml* = 2, 3, 4) or 1,000 times (*ml* = 5, 6) with different random seeds. All standard errors are less than 1%.

The difference in terms of average cooperation rate between the complete representation of the state of the world (D1&D2) and the two incomplete representations of the state of the world (D1, and D2) is clear and it becomes larger the greater the value of memory length *ml* is. When both the player's own decisions and the other player's decisions form the perceived state of the world (D1&D2) the average cooperation rate is much higher than in the other cases.

As we saw in Table 5-1, except in the trivial case where $AT \leq S$, players will never cooperate again in a given state of the world after having received the *Sucker* payoff in that state. When using either of the two incomplete perceptions of the state of the world, there are sets of situations that are represented by the same perceived state of the world for one player but by different perceived states of the world for the other. The size of such sets of situations increases as the memory length *ml* increases. In these sets of situations, one of the players will make several decisions at random in situations which they perceive as novel, but which are represented by one single perceived state of the world for the other player.

This fact strongly increases the chances of the latter player getting a *Sucker* payoff and therefore not achieving a cooperative outcome.

5.4.2. The Tragedy of the Commons game

Aspiration Thresholds

The TC game is more complex to analyse than the PD since at any given time in the simulation players have not necessarily visited what they perceive as a distinct situation the same number of times[31]. Therefore, in a given time-step some players may be making decisions at random while some others may not. This means that we cannot build a table like Table 5-1 for the TC game.

Figure 5-5 shows the results obtained in the TC game when played by 10 players with memory length ml = 1, for different values of M (maximum number of defectors for which the reward is given). Similar results have been obtained when the game is played by 5 and by 25 players.

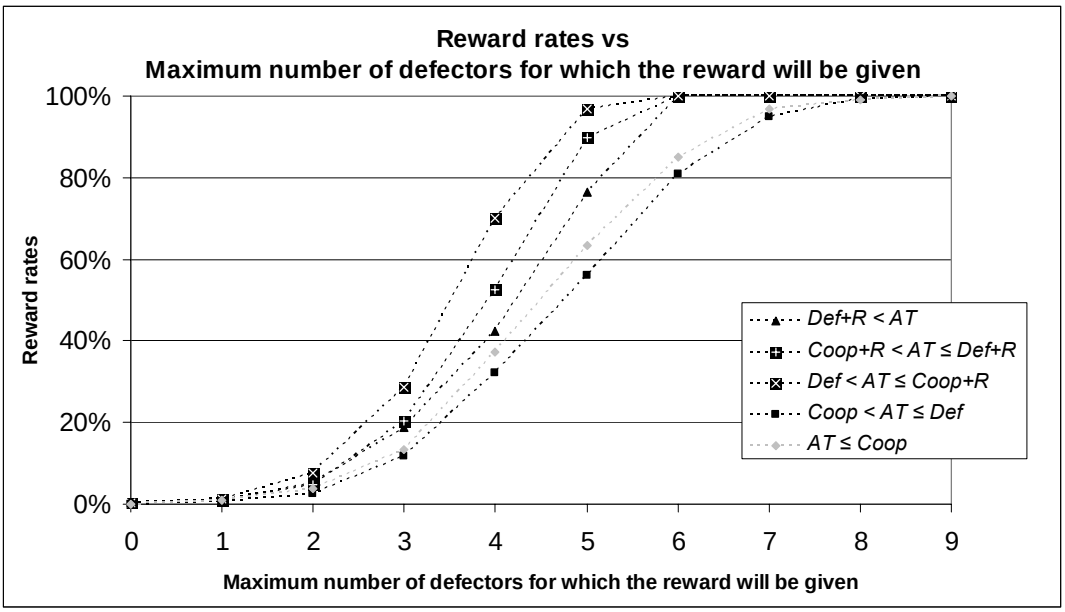

Figure 5-5. Average reward rates for different values of M in the Tragedy of the Commons game played by 10 Players with Memory Length ml = 1. Each represented value has been estimated by running the model 1,000 times. All standard errors are less than 1.5%.

[31] Recall that players know only whether they cooperated or defected, and *how many* others defected. In the TC game, the information provided to the players is thus not complete in the sense that they cannot identify who is defecting, as they could in the PD (since there was only one other player).

Figure 5-5 shows that levels of cooperation strongly depend on the maximum number of defectors for which the reward is given (M). When the requirement is too demanding (low values of M), levels of cooperation tend to be low and the reward is not usually given. On the other hand, for moderate and high values of M ($M \geq 6$), the reward is almost always given[32]. If players have aspiration thresholds greater than *Def-P* then the reward will be given more often than if they choose at random ($AT \leq Coop\text{-}P$). The highest levels of cooperation are achieved when the aspiration thresholds are just above *Def-P*. Levels of cooperation then decrease as aspiration thresholds separate from the optimal value.

Importance of a common perception of the state of the world

To test the importance of a common perception of the state of the world, we put our players on a toroidal 2x5 grid so they could only observe their most immediate five neighbours in their Moore neighbourhood of radius 1. Results are shown in Figure 5-6.

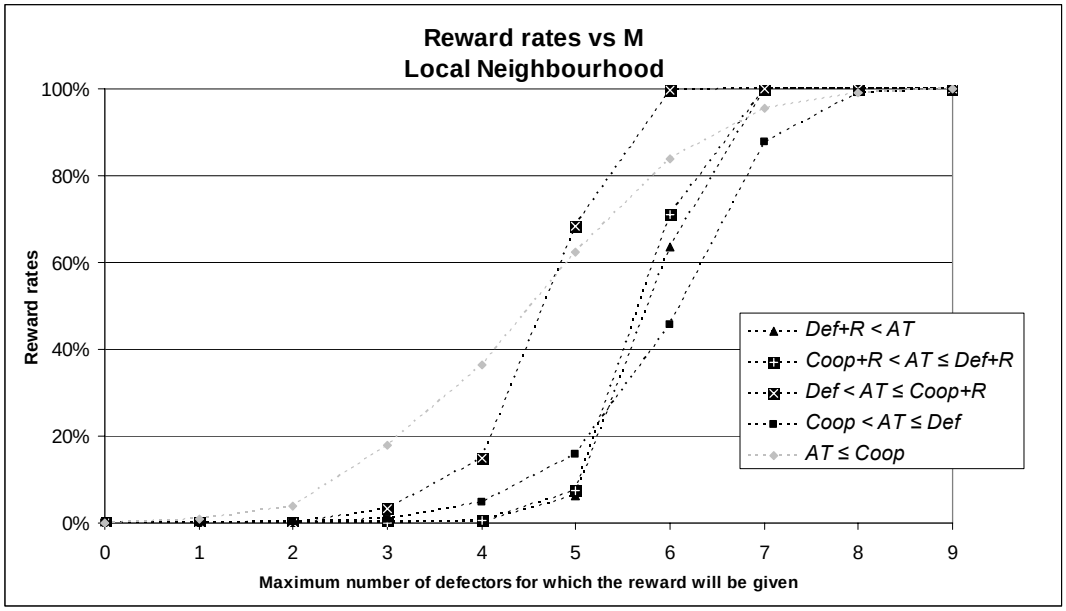

Figure 5-6. Average reward rates for different values of M in the Tragedy of the Commons game played by 10 Players with Memory Length ml = 1. Every player A can observe other 5 players only, who are the only ones that can observe player A. Each represented value has been estimated by running the model 1,000 times. All standard errors are less than 1.5%.

[32] When the game is played by 25 Agents, average reward rates are greater than 80% if $M \geq 15$ and greater than 99% if $M \geq 19$, for any aspiration threshold.

When players can observe only their local neighbourhood, the range of values of *M* to which the reward rate is sensitive is shifted and squeezed to the right. The use of local neighbourhoods sharpens the global movement from defection to cooperation. When players can only observe their neighbours, their global response to changes in the reward programme (parameterised by *M*) is not smooth anymore. Instead, the global behaviour is now better characterised by a hard threshold whose particular value depends on the aspiration threshold of the players forming the society. When players can only observe their neighbours there is a very narrow range of values for *M* where a very small change can make a huge difference.

As in the previous case, the highest levels of cooperation are achieved when the aspiration thresholds are just above *Def-P*. It is once again clear from these results that in CBR, not only *what* is learnt is important, but also *how* it is learnt, and that aspiration thresholds play a crucial role in that process.

5.5. Discussion of the results obtained with the CBR model

The experiments conducted with the CBR model show that cooperation can emerge from the interaction of selfish and unforgiving (but satisficing) case-based reasoners. We are aware that the assumption that Agents make their decisions at random when confronted with a new situation is difficult to maintain. However, Table 5-1 shows that when $AT > Maximin$[13], any positive correlation between the random decisions taken by the Agents will tend to increase levels of cooperation. Similarly, we would expect negative correlations to lead to less cooperative outcomes. The experiments have also shown that the optimal value of the aspiration threshold is just above *Maximin*, and that sharing a common perception of the state of the world strongly increases levels of cooperation.

More importantly, the experiments conducted have revealed a concept of equilibrium which is more relevant than the Nash equilibrium for repeated games played by case-based reasoners: *strictly undominated outcomes* (or individually-rational outcomes). The concept of strictly undominated outcome is defined for one single stage of any game. Its defining property is that no player can be

guaranteed a higher payoff by changing their decision[33] (*i.e.* every player is getting at least their *Maximin*). The concept of strictly undominated outcome is weaker (*i.e*. less restrictive) than the Nash equilibrium: A Nash equilibrium is always a strictly undominated outcome but the reverse is not necessarily true. In particular, in the one-shot PD, bilateral cooperation is a strictly undominated outcome while it is not a Nash equilibrium.

As opposed to the concept of Nash equilibrium (which makes the assumption that the other players will keep their strategies unchanged), the concept of strictly undominated outcome accounts for every possible action that the other players might take. A strictly undominated outcome as equilibrium concept is best defined by negation: if a certain player perceives that by changing their strategy they will always get a higher payoff no matter the other players' response, then the player has a clear incentive to deviate from that outcome, so that outcome cannot be an equilibrium (it is strictly dominated by other outcomes). If, on the contrary, no player has such incentive, the outcome could be an equilibrium. It comes as no surprise that this equilibrium concept is based on outcomes rather than strategies, since case-based reasoners place the emphasis on the case rather than on the rule.

In the PD, the only strictly undominated outcomes are the two bilateral decisions. In the TC the only strictly undominated outcome in which the reward is not given is universal defection; all the outcomes in which the reward is given are strictly undominated.

It can be mathematically shown that all the non-trivial simulations (i.e. those where aspiration thresholds are above the lowest payoff) reported in this chapter must end up in cycles made up of strictly undominated outcomes (Izquierdo et al.,

[33] A slightly more restrictive concept is that of an undominated outcome, in which no player can be guaranteed the same or a higher payoff by changing their decision. The concept of undominated outcome as equilibrium implies that players deviate from an outcome only if it is certain that they will not be worse off by doing so, whereas the *strictly* undominated concept implies that players move away from an outcome only if it is certain that they will be better off by doing so. The concept of undominated outcome as equilibrium is neither weaker nor stronger than the Nash equilibrium.

2004). As we have seen in the previous section, the actual selection among different strictly undominated outcomes can be strongly path-dependent and depends on the specific type of CBR algorithm that players use.

If their aspiration threshold is high enough, players in the CBR model will not accept outcomes in which they are guaranteed a higher payoff by changing their decision once their learning process is finished. However, they are quite naive in the sense that they are not able to infer that the game has locked in to a persistent cycle. In other words, they are not able to infer that the other players will not accept outcomes where they are not getting their *Maximin* either. We can conjecture what would happen if the players were sophisticated enough as to infer, through repeated interaction and learning, the *fact* that the rest of the players are also non-exploitable (*i.e.* they do not accept outcomes where they get a payoff lower than *Maximin*). Assuming (or learning) that the rest of the players are not exploitable can then enable a player X to infer that certain outcomes which give payoffs higher than *Maximin* to this player X will not be sustainable (because they do not yield payoffs higher than *Maximin* to some other player). This inference can make an outcome which was not initially strictly dominated in effect be dominated. In other words, the concept of strict dominance can be applied to outcomes *iteratively* just as it is applied *iteratively* to strategies.

As an example, we have seen that players with a high enough aspiration threshold who play the PD will end up in a cycle made up of bilateral cooperations and/or bilateral defections (the only two strictly undominated outcomes; see Figure 5-7b). If through repeated interaction the players were able to infer that the game will not have any other outcome (because one of the players will not accept it), then they could eliminate the unilateral outcomes from their analysis and apply the concept of outcome dominance for the second time to the (two) remaining possible outcomes. For this to happen, it would have to be mutual belief[34] that the opponent is not exploitable either. When only bilateral decisions are confronted,

[34] A proposition *A* is mutual belief among a set of players if each player believes that *A*. Mutual belief by itself implies nothing about what, if any, beliefs anyone attributes to anyone else (Vanderschraaf and Sillari, 2007).

the only strictly undominated outcome is bilateral cooperation (see Figure 5-7c). When confronted with bilateral cooperation as the only alternative, bilateral defection is not a strictly undominated outcome anymore, since the two players are guaranteed a higher payoff by changing their decision. In other words, bilateral cooperation is the only outcome that survives two steps of outcome dominance in the PD. In the TC game all the outcomes in which the reward is given survive two steps of outcome dominance, and they are the only outcomes that do so. It can be shown that in any game, after applying any number of steps of outcome dominance, the remaining outcomes are not Pareto-dominated by any of the outcomes which have been eliminated.

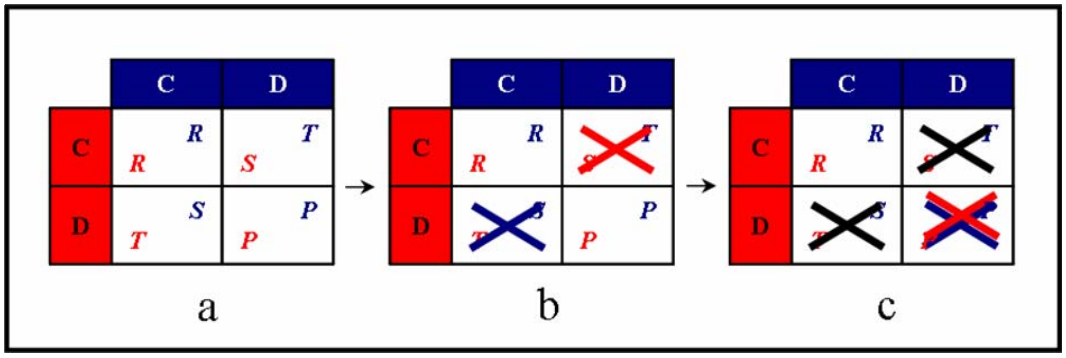

Figure 5-7. Elimination of dominated outcomes. Figure b shows the remaining outcomes after having applied one step of outcome dominance. Figure c shows the remaining outcomes after having applied two steps of outcome dominance. Red crosses represent outcomes which are unacceptable for player Red (row), blue crosses represent outcomes which are unacceptable for player Blue (column), and black crosses represent outcomes eliminated in previous steps.

How players would be able to move from bilateral defection to bilateral cooperation, if indeed they were, is not clear and is a matter for further research. We conjecture that this could be achieved by signalling processes to promote cooperation, or it could emerge from a form of learning by induction, since once the simulation has locked in to a cycle, it does show a general rule or pattern (players get a higher payoff when they cooperate than when they defect). Perhaps induction would then be produced by the simple forgetting of an episode's details and the consequent blurring together in memory of that episode with other similar episodes (Reisberg, 1999). In any case, the movement from bilateral defection to bilateral cooperation would require a non-trivial degree of coordination.

We have seen that if CBR players have a high enough aspiration threshold they are not exploitable in the sense that they do not accept outcomes where they are not getting at least *Maximin*. We find that a more useful definition of rationality in games is that of non-systemic-exploitability. Rational players are not systemically exploitable. According to this definition, cooperation emerges among selfish rational players as soon as it becomes mutual (not necessarily common) belief that the game is being played among rational players. Using Macy's words, cooperation would then emerge among self-interested agents "not from the shadow of the future but from the *lessons of the past*" (Macy, 1998).

5.6. Trembling hands process: the N-CBR model

While useful as a "tool to think with", the CBR model is admittedly rather unrealistic in the sense that simulations end up necessarily with players locked in to a persistent cycle. In this section we consider an extension of the CBR model where players may suffer from "trembling hands" (Selten, 1975) –i.e. they occasionally experiment (or make mistakes) with small probability. Importantly, we also significantly relax the assumptions made about what defines a perceived state of the world and about the decision-making algorithm used by players. These changes make the model more general, slightly more realistic, and the introduction of noise allows us to make more specific predictions. In particular, as in chapter 4, we will characterise the set of outcomes where the system spends a significant proportion of time in the long-term when players experiment with very low probability, i.e. the set stochastically stable outcomes. Such a set of outcomes is a subset of the set of outcomes that can be observed in the model without experimentation. As an example, we will see that in the prisoner's dilemma, mutual cooperation belongs to the latter set but not to the former.

The definition of a case is substantially more general in the *noisy* CBR model (henceforth N-CBR) than in the CBR model. A case (an experience) lived by player i in the N-CBR model comprises:

- The time-step t when the case occurred.
- The *perceived* state of the world at the beginning of time-step t, which is determined by a subset of the decisions undertaken by every player in the game (potentially all decisions by all players, including the case-holder i) in

the preceding ml_i (for *memory length*) time-steps. (Note that different players may have different memory lengths.) When comparing the N-CBR model with the CBR-model it will be assumed that players in the N-CBR model build their perceived state of the world as in the CBR model (see section 5.3).

- The decision made by the case-holder in that situation, in time-step t, having observed the state of the world in that same time-step.
- The payoff that the case-holder obtained after having decided in time-step t.

As in the CBR model, players in the N-CBR model decide what action to select by retrieving the most recent case which occurred in a *similar* situation for each one of the actions available to them. This set of cases, which is potentially empty, is denoted C_i. A case is perceived by the player to have occurred in a *similar* situation if and only if its state of the world is a perfect match with the current state of the world observed by the case-holder. The definition of the decision-making algorithm in the N-CBR model is also substantially more general than in the CBR model. In a certain situation (i.e. for a given perceived state of the world) any particular player i will face one of two possibilities:

- Not every action available to player i is represented in C_i. Given the fact that players in the N-CBR model suffer from trembling hands (this is explained in detail below), this is a temporary situation. No assumptions are made in the N-CBR about how players make decisions in this situation. When comparing the N-CBR model with the CBR-model it will be assumed that players in the N-CBR model use, for this situation, the same decision-making algorithm as in the CBR model (see section 5.3).
- Every action available to player i is represented in C_i. As in the CBR model, in this situation player i selects randomly among those actions with the highest payoff obtained in the set C_i.

As mentioned before, we also assume that players suffer from trembling hands: there is some small probability $\varepsilon \cdot \lambda_i \neq 0$ that player i selects her action randomly instead of following the algorithm above. The ratio λ_i / λ_j determines player i's relative tendency to experiment compared with player j's. The factor ε is a general measure of the frequency of experimentation in the whole population of players. The event that i experiments is assumed to be independent of the event that j

experiments for every $i \neq j$. Different players may experiment in different ways, but it is assumed that player i's probability of selecting any action a available to her when experimenting ($q_i(a)$) is non-zero, potentially different for different actions, and independent of time for all i; these conditions can be relaxed to some extent (Young, 1993). This completes the specifications of the N-CBR model.

This chapter will present some mathematical results valid when the overall probability of experimentation ε tends to zero; all such results are independent of λ_i and of the particular way each of the players experiments. When presenting simulation results, it will be assumed that $\lambda_i = 1$ for all i, and that players select one of their actions randomly and without any bias when experimenting.

5.7. Dynamics of the N-CBR model

The following explains why the N-CBR model has a unique limiting distribution. First, note that any N-CBR model can be formulated as a Markov chain where the state of the system is defined by every player's set of most recent cases that occurred in every possible perceived state of the world for each one of the actions available to her. Given the definition of the set of different states of the world possibly perceived by every player and the nature of the trembling hands noise, it is clear that this Markov chain is finite and has a unique recurrent class (where all actions available to each player i are represented in the set C_i for every state of the world possibly perceived by i). The trembling hands noise guarantees that it is possible to go from any recurrent state to any other recurrent state in a finite number of steps. This basically means that the N-CBR model can be formulated as a uni-reducible Markov chain, which has a unique limiting distribution (Janssen and Manca, 2006, Corollary 5.2, pg. 117).

Thus, note that both the CBR and the N-CBR model can be formulated as finite-state discrete-time Markov chains, but there is a crucial difference between them: the CBR model will end up in one of many possible cycles (the period of some of these cycles is potentially equal to one), whereas the N-CBR process has one unique limiting distribution. Thus, when players suffer from trembling hands, the indefinite cycles where players were locked in the CBR model are broken, and outcomes that occurred infinitely often in the CBR process (like mutual

cooperation in the Prisoner's Dilemma) turn out not to be robust to small trembles. In the following two sections we study the transient and the asymptotic behaviour of the N-CBR process.

5.7.1. Transient dynamics

To explore the transient dynamics of the N-CBR model we focus on the particular N-CBR process merely consisting of adding noise to the CBR model, and we study the Prisoner's Dilemma (PD). As one would expect, the short-term dynamics of this N-CBR process –i.e. when only a few trembles have taken place– are initially similar to the dynamics of the CBR process. How many "a few trembles" are depends on the players' memory and aspiration thresholds; how quickly those "few trembles" occur depends on the probability of trembles happening. Figure 5-8 shows the proportion of outcomes where both players are cooperating (cooperation rate) in the PD for different values of both players' memory $ml_i = ml$ and aspiration threshold AT, and for different values of the overall probability of trembles ε. The cooperation rates shown in Figure 5-8 are calculated over time-steps 1001 to 1100.

A word of caution about Figure 5-8 is that, because it shows the data collected at a predetermined range of time-steps (1001–1100), it represents the short-term behaviour of those series for which 1000 time-steps are not enough to approach their long-term behaviour (e.g. $ml_i = 5$) but, on the other hand, it represents the long-run behaviour for some other series (e.g. those series for which 1000 time-steps are enough to reach it, like series with $ml_i = 0$, and $\varepsilon \neq 0.001$).

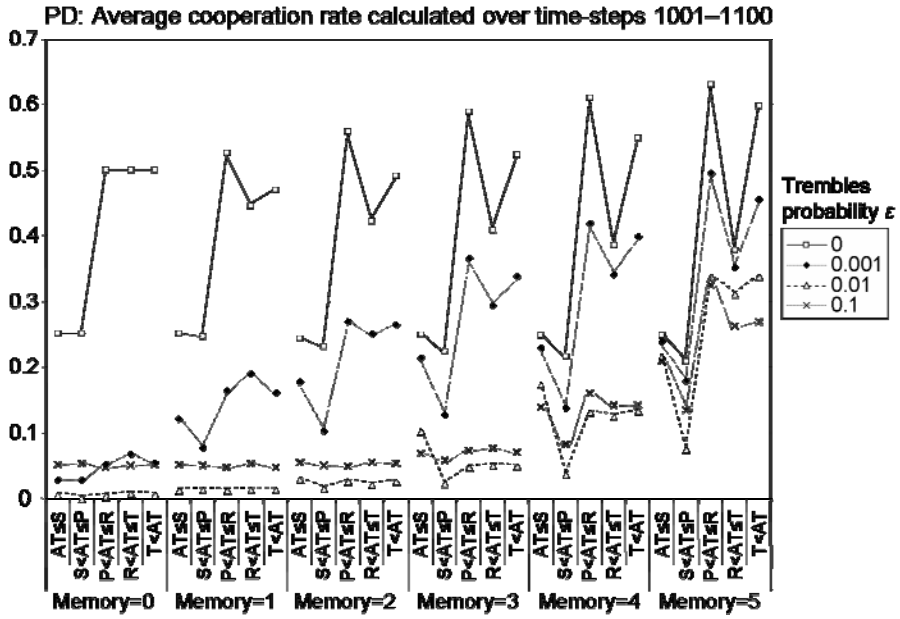

Figure 5-8**.** Average proportion of outcomes where both players are cooperating in the Prisoner's Dilemma (PD), calculated over 100 time-steps starting at time-step 1001, and using 500 simulation runs for each data point. The payoffs in the game are represented by its initial letter: S for Suckers, P for Punishment, R for Reward, and T for Temptation.

5.7.2. Asymptotic behaviour

Once enough trembles have taken place in every situation distinctively perceived by any player, the dynamics of the N-CBR model approach its asymptotic behaviour. The following proposition shows that a very broad range of N-CBR models share the same asymptotic behaviour:

Proposition 5-1: Assuming that every player has a common perception of the state of the world[35], the asymptotic behaviour of the N-CBR process is independent:

1. of the specific structure of the perceived state of the world (i.e. the algorithm used to construct it), and
2. of the decision-making algorithm employed by each player *i* when she has not explored every action available to her in a *similar* situation (i.e. when not every action available to player *i* is represented in C_i).

[35] This means that any two situations that look the same to one player will also look the same to every other player and any two situations that look different to one player will also look different to every other player.

Proposition 5-1, which is proved in Appendix B, implies that the asymptotic dynamics of all the simulations shown in Figure 5-8 are independent of the players' memory (see point 1 in the proposition) and of their aspiration thresholds (see point 2 in the proposition). Thus, for example, the long-run cooperation rate in the PD (calculated analytically) is $4.985 \cdot 10^{-2}$ for $\varepsilon = 0.1$, $4.978 \cdot 10^{-3}$ for $\varepsilon = 0.01$, and $4.998 \cdot 10^{-4}$ for $\varepsilon = 0.001$. As we can see in Figure 5-8, the series with low memory ($ml_i = 0$ or $ml_i = 1$) and high probability of trembles ($\varepsilon = 0.1$ or $\varepsilon = 0.01$) quickly converge to their limiting values; for those parameterisations 1000 time-steps are sufficient to reach the long-run behaviour of the process. If we represented the data in Figure 5-8 after a sufficiently high number of time-steps, the value of every data point with $\varepsilon \neq 0$ would only depend on the probability of trembles ε (and on λ_i and $q_i(\cdot)$ generally), and it would approach the analytically calculated values presented above (calculated for $\lambda_i = 1$, and $q_i(\cdot)$ unbiased). Something which is clear in Figure 5-8 is that whereas mutual cooperation usually forms part of the cycles in the CBR model, it cannot be sustained in the long-term when small trembles occur.

To summarise, the dynamics of the N-CBR model follow a transition from a very path-dependent distribution similar to that corresponding to the CBR model, to a very different distribution which is only dependent on the probabilities with which trembles occur.

5.7.3. Stochastic stability

Having seen that the asymptotic behaviour of the N-CBR model is only dependent on the structure of trembles (assuming a common perception of the state of the world), a natural question is: What outcomes can be observed with probability bounded away from zero in the long-run as the probability of trembles ε tends to zero? Following Young (1993), such outcomes will be called *stochastically stable*. It turns out that whether an outcome is stochastically stable or not is independent of λ_i and of $q_i(\cdot)$ (Young, 1993).

Young (1993) provides a general method to identify stochastically stable *states* in a wide range of models by solving a series of shortest path problems in a graph. In our model there are more states than outcomes, but identifying stochastically

stable outcomes when the set of stochastically stable states is known is straightforward. Young's method uncovers an important feature of stochastic stability: stochastic stability selects states which are easiest to flow into from *all* possible states of the system. This contrasts with most notions of equilibrium based on full rationality. As Young (1993) notes, risk dominance "selects the equilibrium that is easiest to flow from every other equilibrium considered in isolation". Similarly, Nash stability is determined only by unilateral deviations from the equilibrium.

In this section we present some features to identify stochastically stable outcomes when reasoning is based on singletons of distinct prior outcomes. We start with a necessary condition for outcomes to be stochastically stable in N-CBR models (it is not assumed that players must share a common perception of the state of the world).

Proposition 5-2: In all N-CBR models, every stochastically stable outcome is individually rational.

The proof of Proposition 5-2 can be found in appendix B. Proposition 5-2 is a useful necessary condition to identify outcomes which cannot be stochastically stable but, except in very simple games (e.g. see Figure 5-9A), it is not sufficient to characterise the set of stochastically stable outcomes. To try to identify features that make outcomes stochastically stable we developed a computer program in Mathematica© that calculates the exact long-run probability that any 2-player game is in each possible outcome when the probability of trembles tends to zero. To calculate such probabilities, we did have to assume that players share a common perception of the state of the world. Using the computer program, we came to the following conclusions:

- Stochastically stable outcomes are not necessarily Nash equilibria (e.g. see the game of Chicken in Figure 5-9B).
- In fact, some players in some stochastically stable outcomes may be choosing strictly dominated strategies (e.g. see the game represented in Figure 5-9C).

- Nash equilibria are not necessarily stochastically stable (e.g. see the game of Stag Hunt in Figure 5-9D).
- Stochastically stable outcomes can be Pareto dominated by outcomes which are not stochastically stable (e.g. see the Prisoner's Dilemma game in Figure 5-9E). However, it can be proved that stochastically stable outcomes cannot be Pareto dominated by outcomes which are one tremble away and which are not stochastically stable. Thus, in the game represented in Figure 5-9C, for example, if we knew that outcome (3,3) is stochastically stable, then we could infer that (4,4) would have to be stochastically stable too.
- Stochastically stable outcomes can Pareto dominate outcomes which are not stochastically stable (e.g. see game represented in Figure 5-9A).

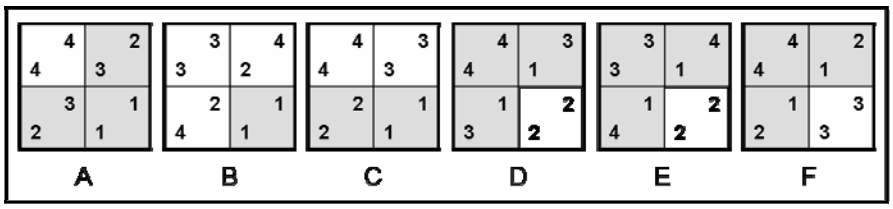

Figure 5-9. Stochastically stable outcomes (highlighted in white) in various 2-player 2-strategy games. Payoffs are numeric for the sake of clarity, but only their relative order for each player is relevant.

Intuitively, note that trembles can destabilise outcomes in two different ways: by giving the deviator a higher (or equal) payoff, or by giving any of the non-deviators a lower payoff[36]. The first possibility is related to the concept of Nash equilibrium, whilst the second is related to the concept of "protection" (Bendor et al., 2001b). As explained in section 4.7 when studying the Bush-Mosteller learning algorithm, an outcome is protected if unilateral deviations by any player do not hurt any of the other players. Bendor et al. (2001b) show that under a very wide range of conditions, reinforcement learning converges to individually rational outcomes which are either Pareto optimal or a protected Nash

[36] Non-deviators could get a lower payoff after a tremble and still keep choosing the same action if the payoff obtained when the tremble occurs is higher than any of the payoffs that the non-deviator obtained when she last selected each of the other possible actions.

equilibrium. The same is not true for the model we study in this chapter (see the game represented in Figure 5-9F), but protected strict Nash equilibria are very relevant here too (as they were proved to be in the Bush-Mosteller model too; see section 4.7): if there is a protected strict Nash equilibrium in a game, then there is at least one state which is robust to any one single tremble, and the outcome that follows such a state in the absence of trembles is the protected strict Nash equilibrium. In fact, it can be shown that the only stochastically stable outcome in any 2-player 2-strategy game with a (necessarily unique) protected strict Nash equilibrium is such equilibrium. The extension of this result to more general games is left for future work.

5.8. Conclusions of this chapter

This chapter has explored the implications in strategic contexts of reasoning by single and distinctive past experiences as opposed to reasoning by abstract rules (strategies). While the short-term dynamics of models where players base their decisions on past experiences are very dependent on the specifics of such models, a very wide range of models behave similarly in the long-term. In particular, a large collection of models where players experiment from time to time share the same set of stochastically stable outcomes (outcomes that persist in the long-run when trembles are very rare).

Stochastically stable outcomes are necessarily individually rational, but a clear relationship between them and Nash equilibria, or Pareto optimality, has not been found. Nash equilibria may, or may not, be stochastically stable, and stochastically stable outcomes may, or may not, be Nash equilibria. The same applies for Pareto optimal outcomes. A concept that is indeed closely related to stochastic stability is the concept of protected strict Nash equilibrium. In particular, in 2-player 2-strategy games with a protected strict Nash equilibrium (which is necessarily unique), the only stochastically stable is such an equilibrium. The importance of the impact of unilateral deviations on non-deviators for the stability of outcomes was also highlighted in chapter 4. This seems to be a recurring observation in learning game theory: if a unilateral deviation harms another player, the non-deviator who has been hurt may choose to select a different strategy in the subsequent period, thus compromising the

stability of the original strategy profile. A unilateral deviation that does not hurt any non-deviator is less likely to trigger a change of strategy in the non-deviators.

In broader terms, this chapter has proposed a new algorithm to narrow the set of expected outcomes in games. This method, i.e. iterative elimination of dominated outcomes, is a logical process through which outcome-based reasoners can arrive at sensible (*i.e*. Pareto optimal) outcomes in games. The only outcome that survives two steps of iterative elimination of dominated outcomes in the Prisoner's Dilemma is mutual cooperation. Thus, this chapter has shown that reasoning by outcomes leads to solution concepts significantly different from those present in the classical game theory literature (where reasoning is conducted using strategies as the key concept). Interestingly, one could argue that there is no a priori logical argument why rationality in game theory should be defined in terms of strategies rather than outcomes. Players in game theory do select a strategy (rather than an outcome), but the payoff they receive (i.e. their measure of performance) is determined by the resulting outcome, which is only partially determined by their selection of strategy. Thus, when defining rationality in game theory, it seems as natural to define it in terms of outcomes as the key concept (i.e. rational players do not choose dominated outcomes), as to define it using strategies (i.e. rational players do not accept dominated strategies). Reasoning by outcomes may even be a more natural way of modelling real human behaviour. Admittedly, the definition of rationality by outcomes proposed here implies some dynamicity (note the sentence: "players *do not accept* dominated outcomes"), whereas the definition of dominance reasoning does not. However, it is also true that, as explained in section 2.2.2, the concept of dominance reasoning is hardly ever enough to narrow the set of expected outcomes in games significantly, and when stronger concepts of rationality based on strategies are brought into play, issues at least as worrying as those that may be raised when defining outcome-based rationality often appear. These issues will be discussed further in chapter 7.