# A Comparison between Cognitive and AI Models of Blackjack Strategy Learning

Marvin R.G. Schiller and Fernand R. Gobet

Brunel University, London, UK
{marvin.schiller,fernand.gobet}@brunel.ac.uk

**Abstract.** Cognitive models of blackjack playing are presented and investigated. Blackjack playing is considered a useful test case for theories on human learning. Curiously, despite the existence of a relatively simple, well-known and optimal strategy for blackjack, empirical studies have found that casino players play quite differently from that strategy. The computational models presented here attempt to explain this result by modelling blackjack playing using the cognitive architecture CHREST. Two approaches to modeling are investigated and compared; (i) the combination of classical and operant conditioning, as studied in psychology, and (ii) SARSA, as studied in AI.

## 1 Introduction

Research in AI and cognitive science has made important contributions to understanding the difficulties underlying a myriad of learning tasks by devising and investigating learning algorithms. In this paper, we address the question of how human learning, which is governed by underlying psychological mechanisms, is modelled and investigated using a cognitive architecture (in our case, CHREST [1]). As the learning task, we use a game that relies both on chance and strategy, and for which empirical data shows that human players deviate from theoretically optimal strategies: blackjack. Models in CHREST simulate the information-processing of human players; they play the game, observe the outcomes, process and store the relationships between blackjack hands, actions and outcomes in long-term memory, and select actions accordingly. CHREST uses a mechanism that implements emotional memory, i.e. patterns of information in memory (chunks) may be associated with emotional tags, which are learned from experience with the environment. Using this basic framework, we investigate models based on theories in psychology and decision-making, and study how they compare to SARSA (cf. e.g. [2]), an AI algorithm modelling reinforcement learning, in explaining data from human casino players.

This paper is organised as follows. In Sect. 2 we discuss previous work investigating strategies in blackjack, both in psychology and AI. In Sect. 3 the cognitive architecture CHREST is introduced. Sect. 4 describes our modelling and Sect. 5 presents the results, which are discussed in Sect. 6.

B. Glimm and A. Krüger (Eds.): KI 2012, LNCS 7526, pp. 143–155, 2012.

## 2 Blackjack Strategies: Modelling and the Role of Learning

Blackjack is played by one or several players (independently) against a dealer. At the start of a game, each player makes a bet and is dealt two cards face-up and their combined value is considered (figure cards count as 10 except the ace, which may count as 1 or 11, to the favour of the player). The dealer also obtains two cards, one of which is dealt face-up (the upcard). The goal of the player is to obtain a total that is as close to 21 as possible without exceeding 21, in which case the player immediately loses the game and thus the bet (going bust). If the initial two cards of the player add to 21 (blackjack), the player immediately wins the game and 2.5 times the amount of the original bet. Such hands that include an ace that may count as 11 without the total exceeding 21 are called soft (as opposed to hard) hands. Players may successively request to be dealt additional cards (hit) or to content themselves with the current total (stand). Further actions the player may take are splitting a pair hand, and doubling a bet (see e.g. [3, Sect. 3]), but these actions are not relevant for the remainder of this paper. After all players have made their choices, it is the dealer's turn to play according to a fixed rule (which may vary in different casinos); usually the dealer is required to hit at a total of 16 or less and to stand at 17. If the dealer busts, all players who have not bust receive 2 times the amount of the original bet (i.e. a net win of the size of the bet). If the dealer stands at a score of at most 21, the remaining players with a higher score win 2 times the original bet, otherwise they either draw (and receive back their bet) or lose their bet.

Wagenaar [4] notes that – even though strategies exist to maximise the returns of playing – the performance of blackjack players in the casino is not optimal. The question of what strategies people actually adopt when playing blackjack, and how they relate to learning, has not been answered conclusively. In this paper we propose a cognitive model (implemented using the cognitive architecture CHREST [1]) that relates blackjack strategy to theories in psychology and decision-making, and compare this to a traditional AI algorithm, SARSA.

### 2.1 Blackjack Strategies

*Never bust* is a strategy where the player hits at a total of 11 and below, and stands at a hard total of 12 or more. According to simulations [4], players using *never bust* are expected to lose 8% of their original investment on average. *Mimic the dealer* is a strategy where the player hits at 16 and stands at 17, like the dealer. The expected loss is 6% per game according to [4]. Both of these strategies are inferior to the *basic strategy* introduced by [5]. It can be represented in the form of decision tables taking into account the player's total and the dealer's upcard, and which prescribe one out of four actions (stand, hit, double, split). Different tables apply to hard hands, soft hands, and pair hands. Wagenaar [4] notes that this strategy can be learned very easily. It results in an expected loss of 0.4 % per game, which is relatively favourable for the player. In combination with a supplementary technique (card counting, as discussed by [4]), a positive expected return can be achieved. Sub-optimal play results in a larger house edge, i.e. expected losses of the player. Walker et al. [6] found that Australian players violated the basic strategy on 14.6% of the hands, resulting in a house edge of 2.4 % (instead of 0.8% with the basic strategy[1]).

To assess what type of strategy blackjack players in the casino are actually playing, Wagenaar studied the games[2] of 112 players and compared their play to the basic strategy. For each combination of the player's total and the dealer's upcard, it was established how often the players (on average) deviated from the basic strategy. Table 1 is adapted from [4] and shows the proportion of deviations from the basic strategy for hard non-pair hands (pair hands, which allow for splitting, were not further investigated in [4]). The table includes only players' totals from 12 to 17, since players always hit at 11 or less, and always stood at 18 or more. Wagenaar found that players are more likely to violate basic when they are required to hit (the underlined area) than when they are required to stand. Wagenaar discusses possible reasons for this kind of "conservatism" of standing where hitting offers a greater chance to win (e.g. regret minimisation, delaying bad news, blaming the dealer's luck). Wagenaar formulates this bias in the players' actions as a linear logistic model, but does not explore whether or how it might be related to learning.

### 2.2 Modelling Learning in Blackjack

A number of studies on blackjack strategies and learning have been contributed by researchers in mathematics and AI, who were either interested in optimal playing strategies or efficient machine-learning algorithms. Less is known about the blackjack skills of actual players, as investigated in [4], [6] and [7].

Work by mathematicians and statisticians (e.g. [5]) makes use of knowledge of the mechanics of blackjack (composition of decks, random drawing) in their search for optimality. In contrast, strategy acquisition in blackjack can be considered as a learning problem based solely on playing experience, which makes the analysis more challenging, and which is done in AI to investigate the capabilities of machine learning algorithms (e.g. bootstrapping [8] and evolutionary algorithms [3]). The work by Perez-Uribe and Sanchez [2] is interesting in that they use the SARSA algorithm, a reinforcement learning mechanism based on temporal difference learning and Q-learning. However, this work does not analyse the relationship between their models and the behaviour of human players. Furthermore, this work did not consider the value of the face-up dealer's card, and therefore it does not offer the possibility to compare the learned strategies to the basic strategy, one of the objectives of this paper.

Reinforcement learning in the context of decision-making tasks has been intensively studied with the Iowa Gambling Task, a kind of four-armed bandit problem (cf. [9]). Subjects have to select cards from four decks with different

[1] Differences wrt. the house edge reported in [4] are due to rule variations.

[2] Since observers had no control over the length of stay, the number of hands recorded for each player varies (median=74 hands, cf. [4]).

schedules of rewards and punishments. [10] and [9] propose an elaborate model for subjects' decision making behaviour on the task, the expectancy valence (EV) model. It models the learning of players' expectations and action selection via softmax selection/Boltzmann exploration. In this paper, we extend the use of some of these principles to the study of blackjack play.

## 3 CHREST

CHREST (Chunk Hierarchy and REtrieval STructures) is a cognitive architecture that enables the modelling of human processes of perception (in particular visual attention), learning and memory. CHREST is a symbolic architecture based on chunking theory [11] and template theory [12]. Chunking theory posits that information is processed, learned and retrieved in the form of patterns, which can be used as one coherent unit of knowledge, and which are referred to as chunks. CHREST is composed of an input analysis component, a short-term memory component (for different modalities: visual, auditory, action) and a long-term memory component which is organised in a network structure. Technically, patterns form nodes in the network structure of long-term memory. Retrieval is via an index structure referred to as discrimination network, which is learned incrementally. Furthermore, cross-links can be learned within long-term memory within and across chunks of different modalities. Patterns are formed when information is perceived via the input analysis component and passed on to short-term and long-term memory, where they are learned – i.e. integrated into the network structure – incrementally. Any chunk in long-term memory (LTM) can be associated with an emotional tag that is retrieved when the chunk is retrieved. In general, emotional tags in CHREST follow the paradigm of [13] and [14] by representing emotions as combinations on several dimensions of primary basic emotions (e.g. joy, acceptance, fear, surprise, sadness, disgust, anger, anticipation according to [14]). In this paper, however, we only use two dimensions of emotions (joy and sadness), in keeping with the parsimony of similar previous models for the Iowa Gambling Task (cf. Sect. 2.2). Emotional tags are learned via an association learning mechanism using a so-called $\Delta$-rule (illustrated in the next section). This rule is part of psychological theories on classical conditioning [15] as well as the decision-theoretic model proposed by [10].

CHREST runs as a computer program in Java (with interfaces for scripts in other languages), to enable simulations and testing of cognitive models. CHREST has previously been used to model phenomena in various domains of human information-processing and expertise, including board games (Go, chess and awalé), language acquisition in children, and physics.

## 4 Modelling

CHREST models played blackjack, to investigate in how far the learning implemented by these models accounts for (i) the behaviour observed by Wagenaar [4] as described in Sect. 2, and (ii) the choices that the basic strategy prescribes instead. Since the actions of splitting and doubling are not relevant for modelling Wagenaar's data, our model is simplified by only considering hitting vs. standing (like [2]). As a further simplification, each game is dealt from a complete deck of cards. For simplicity we also assume that bets are held constant, and wins and losses are always represented as multiples of "1" bet. Our model is based on the hypothesis that players experience constant reinforcement and reward (of positive and negative valence) while playing, which follows a random ratio schedule. This experience is likely to enter the player's memory and to influence decision-making, in conflict or in addition to fixed strategies. For making an action, the model (i) visually recognises the total of the player's hand, (ii) visually recognises the value of the dealer's upcard, and (iii) retrieves the set of possible actions in action memory associated with that situation. Depending on previous experience, these situation-specific actions are associated with emotional tags of positive and/or negative valence. Depending on these values, the action to be performed is selected, the immediate outcomes are observed, and the model adjusts its expectations (using the $\Delta$-rule).

### 4.1 CHREST Model (Model 1)

The model starts out with a representation of the possible losing and winning outcomes of blackjack. Monetary wins and losses carry emotional tags attached to these outcomes as follows; wins are directly represented in the "joy" dimension and losses in the "sadness" dimension. The possible outcomes are (i) losing (joy:0, sadness:1), (ii) blackjack win (joy:2.5, sadness:1), (iii) ordinary win (joy:2, sadness:1), (iv) push (joy:1, sadness:1). An alternative approach is to map the net wins onto a single dimension of "joy" (which can then also take negative values). The second approach assumes that players mentally offset bets and wins prior to experiencing their rewarding effect. We additionally included this variant of our model in the analysis, as discussed in the results.

Each game of blackjack requires the player to make one or several choices based on the current "state" of the game; i.e. the situation described by the player's cards and the dealer's upcard. Each game is terminated with a win, loss or draw after a finite number of iterations of the following steps.

**1) Recognition**. The model receives its own hand and the dealer's upcard as input. The model retrieves a chunk that represents the combination of its own total, together with an indicator whether the hand is soft or hard, and the dealer's upcard from long-term memory (e.g. "14-10-soft" if presented with Ace, 3 and dealer's upcard 10). If such a chunk does not exist or is incomplete, learning of such a chunk (according to chunking theory, cf. [1]) takes place instead. If the chunk has been linked to actions or carries an emotional tag (due to previous playing), these are retrieved. Any such "player's hand" or "state" chunk may have been linked with two different action chunks (hit/stand), each of which may have an emotional tag (resulting from previous experience with the action for that specific hand).

**2) Action Selection.** Based on the emotional tags for hit/stand actions linked to the chunk representing the player's hand in long-term memory, one is chosen over the other probabilistically via softmax selection as follows. For both options, the expected value of taking that action is taken to be the difference between the "joy" and "sadness" values of their emotional tags (and 0 otherwise). Let $Ev_{\text{STAND}}(x, y, z, t)$ denote the expected value of the STAND action for the hand characterised by a total of x, a dealer's upcard value of y, and the indicator for soft/hard hands z. Furthermore, assume that the model has previously encountered the current choice situation (e.g. "14-10-soft") $t$ times. Then the probability that the model stands (rather than hits) is defined by the Boltzmann softmax:

$$Pr[\text{STAND}(x, y, z, t)] = \frac{exp(\theta(t) \cdot Ev_{\text{STAND}}(x, y, z, t))}{exp(\theta(t) \cdot Ev_{\text{STAND}}(x, y, z, t)) + exp(\theta(t) \cdot Ev_{\text{HIT}}(x, y, z, t))} \quad (1)$$

where $\theta(t) = (\frac{t}{10})^c$. The function $\theta(t)$ regulates the transition of the model from exploration in the beginning of learning (i.e. choosing actions at random with equal probability) towards exploitation of the learned values (where differences in the learned values $Ev$ determine the choices to a large degree), cf. [9]. The parameter $c$ represents the rigour with which the model transits from exploration to exploitation, if $c$ is chosen to be positive (and vice versa otherwise). Thus, when nothing is known about both actions, chances of either being selected are fifty-fifty. With more experience, the model becomes more sensitive to the differences in the emotional tags and makes a more rigorous selection. The value $1/\theta(t)$ is called temperature [9]. This form of action selection is analogous to the use of softmax selection in modelling the Iowa Gambling Task with the EV model [9,10]. A difference, however, is that our model maintains individual temperatures for the different constellations of the player's and dealer's hand, whereas the temperature in the EV model is global. This way, we take into account the inherent imbalance in how often different constellations occur in blackjack.

**3) Action and Reinforcement/Conditioning.** The model carries out the selected action and obtains the results from the environment; either (i) the game ends with a win/loss/draw, or (ii) the model hits and remains in the game, and can thus make another choice. In both cases, association learning takes place, where the hand and the selected action are associated with positive and negative emotions based on the outcome. This uses the $\Delta$-rule, which applies when a chunk $x$ (a hand, or an action) is followed by a reward $r$. For each emotional dimension $e$, the emotional value of the chunk $x$ is updated by the amount

$$\Delta x_e := \alpha(r_e - x_e). \quad (2)$$

The parameter $\alpha$ (with $0 \leq \alpha \leq 1$) is called the learning rate or update rate[3]. In case (i) the reward is a blackjack win/loss outcome as defined at the beginning of

[3] Low values of $\alpha$ represent slow learning and slow forgetting, whereas high values represent a bias towards recent events and a more limited memory [9,10].

this section, then both the (previous) hand and the action are credited this way (classical and operant conditioning). In case (ii), the reward is represented by the emotional tag associated with the new hand. For example, if the HIT action was chosen for the hand "14-10-soft", and the player is dealt a king, the new hand is "14-10-hard". The emotional tag associated with "14-10-soft" is updated with the $\Delta$-rule and "14-10-hard" as a reward. Similarly, the emotional tag of the action chunk"14-10-soft→hit" is updated in the same way. Thus, the values of the emotional tags for hands propagate from those hands that are likely to receive immediate reward (or punishment) towards those that are more likely to represent an intermediate stage of the game. The model continues with the next choice as described by **1)**.

### 4.2 Attribution

In his analysis of casino players, Wagenaar [4] observed that the probability of hitting correlates with the probability of busting after drawing one card. As illustrated in Table 1, players are likely to hit at a hard hand of 12 and to be gradually more likely to stand as totals approach 16. This raises the question whether the behaviour of players can be conceived as "fitting" their decisions to the probability of busting, and ignoring other aspects of decision-making, and several hypotheses for this behaviour (attribution bias, delaying) are discussed by Wagenaar. We assume that affective conditioning requires stimuli to co-occur in short-term memory. This contiguity is playing an important role for crediting the player's actions with the emotional consequences (losing or winning). Perceptions that happen in between the player's action and the outcome (e.g. the dealer's actions), are likely to enter short-term memory and interfere with this contiguity. This results in a bias towards learning actions that are immediately punished or rewarded (busting or having blackjack) and against outcomes that involve the dealer's actions. In this paper, we use an explicit parameter that quantifies this bias, rather than model the perception of the dealer's actions in detail. We define an attribution bias $att \in [0,1]$ that inhibits the learning of outcomes delayed by the dealer's actions. Technically, we use a reduced learning rate $\alpha' = \alpha(1 - att)$.

### 4.3 SARSA (Model 2)

The above described learning mechanism is similar, but not identical to a well-known reinforcement learning algorithm in AI, SARSA. Since SARSA has been described in detail elsewhere (e.g. [2]), we concentrate on the differences to the above algorithm. The simulations in this paper use a modification of Model 1 that implements a version of SARSA. Instead of using the above $\Delta$-rule, actions are reinforced by the rule

$$\Delta x_e := \alpha(r_e + \gamma x'_e - x_e), \tag{3}$$

where $x$ is the emotional tag associated with the action that has been taken, $r$ is the emotional tag associated with the immediate reward (wins/loss/draw if the game has finished, nothing otherwise), and $x'$ is the emotional tag associated with the next action that the model will take (according to action selection) if the model is allowed another choice (in case the game continues after hitting). This is done for all emotional dimensions indexed by $e$, sadness and joy. The parameter $\gamma$ (in the range [0,1]) is used to tune down the contribution of the intermediate reward $x'$ represented by staying in the game vis-a-vis actual reward $r$. Apart from the incremental learning of chunks, which we retain from the previous model, this implements a version of SARSA. In brief, the differences between the two variants of our model are the following:

- Model 1 maintains estimated values for both the player's hands and actions related to hands, Model 2 only for the latter.
- When Model 1 hits and is presented with a new choice, the hit action is credited by the general (action-unspecific) value of the new hand. Instead, Model 2 credits the action with the hypothetical value of the next action that will be chosen (i.e. it looks further ahead).
- The parameter $\gamma$ regulates the contribution of expected vs. actual reward.

### 4.4 Model Fitting

Because the game of blackjack has an important stochastic element, for each of the models described above, and each set of parameters, we constructed ten instances with the same parameters, and calculated the fit relative to (i) the basic strategy, and (ii) the data from [4] in Table 1. Since this procedure is relatively time-consuming, and models have three parameters ($\alpha$, $c$, and either *att* or $\gamma$) we performed a relatively coarse-grained grid search (550 combinations of parameters) to explore different models and to optimise their parameters (using the sum of squared errors of the models' percentages of standing in the different cases). We consider it more important to understand how the different models compare rather than pinpoint the exact location of the optimal parameters (which, due to the stochastic nature of the game, is very difficult anyway). We applied this procedure to models that played 10,000 blackjack games each (equivalent to roughly 194–333 hours of play, according to [4][4]). To ease the analysis, we assume that all models always hit at a total of less than 11, and stand at a hard total of at least 18. It is a plausible assumption that human players start off with a similar rule of thumb when learning blackjack, and Wagenaar's data shows that human players generally do not violate this simple rule.

## 5 Results

We measured in how far our models account for the learning of both (i) observed decision-making by casino players (as represented by Wagenaar's data in Table 1) and (ii) ideal decision making (as represented by the corresponding excerpt of the basic strategy, indicated by underlining in the table) by calculating $r^2$ (for

[4] Wagenaar [4] estimates that 30 hands in succession equal 35-60 minutes of playing.

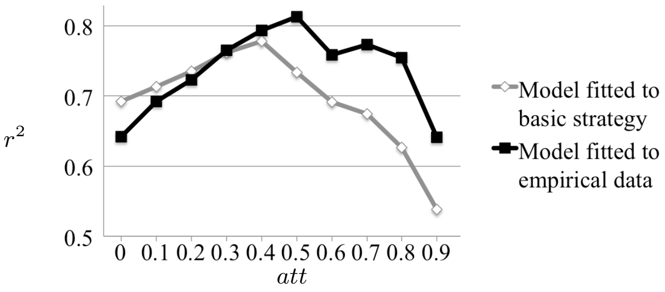

**Fig. 1.** Comparison illustrating the role of *att* for the fit of the CHREST model wrt. the basic strategy ($\alpha = 0.1$, $c = 3$) and Wagenaar's data ($\alpha = 0.2$, $c = 3$)

the percentage of standing, as opposed to hitting, in the different cases) and comparing the patterns in the decision tables. Table 2 shows the decision making pattern of Model 1 fitted to Wagenaar's empirical data, with $r^2 = 0.81$. The main phenomena in Wagenaar's data are present: when the dealer's total is low, players sometimes hit when they should stand. When the dealer's total is high, then the model is more likely to stand the higher the player's total.

Table 3 shows the decision making pattern of Model 1 fitted to the basic strategy. The fit is $r^2 = 0.79$. The model uses an attribution bias of 0.4. This bias is less than for the model fitted to the empirical data (as expected), but it is nevertheless surprising, since one would expect that a bias is likely to be detrimental to learning the optimal strategy. A possible explanation is that the attribution factor makes the model play more conservatively (i.e. stand rather than hit), and since the table contains more situations where the basic strategy mandates standing over hitting, our optimisation is slightly biased towards those situations (and thus, conservatism). Despite the relatively good fit, the model still falls noticeably short of attaining the basic strategy, illustrating how difficult it is to learn the strategy by playing. To assess the contribution of the attribution factor *att*, we compare the fit of the two models while varying *att*, as shown in Fig. 1. This shows that the attribution factor indeed contributes towards the fit of the model for Wagenaar's data. Interestingly, the model does not seem to depend on using separate dimensions of joy and sadness – a similar fit is obtained using only one emotional dimension where rewards reflect the net win only ($r^2 = 0.80$ for Wagenaar's data with $\alpha = 0.4$, $c = 3$, $att = 0.4$ and $r^2 = 0.76$ for the basic strategy, with $\alpha = 0.2$, $c = 5$, $att = 0.3$).

Table 4 shows the decision making pattern using Model 2 with parameters $\alpha = 0.2$, $c = 2$ and $\gamma = 0.1$, which results in an unexpectedly close fit to Wagenaar's data, $r^2 = 0.91$. In particular, this model exceeds the model in Table 2 by better representing the trend of players to hit at a low total and to stand at a high total when the dealer's upcard is high. The fact that $\gamma$ is found to have a very low value means that the contribution of expected reward counts only with a factor of 0.1 relative to immediate reward, and thus represents a strong bias towards immediate reward. By contrast, our grid search produced only moderate results when fitting Model 2 to the basic strategy, not better than $r^2 = 0.6$.

**Table 1.** Percentage of decisions of casino players that violate the basic strategy for hard non-pair hands. Underlined percentages indicate those cases where the basic strategy requires hitting (in all other cases the basic strategy requires standing). Table adapted from [4] and shading was added (gray represents the percentage of hitting).

| | Dealer's Upcard | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Player's total | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 12 | 14.5 | 33.7 | 47.7 | 44.1 | 29.9 | 9.4 | 9.0 | 9.3 | 7.7 | 3.7 |
| 13 | 49.5 | 23.3 | 17.4 | 8.2 | 8.2 | 28.2 | 22.5 | 17.6 | 17.8 | 8.3 |
| 14 | 24.5 | 10.4 | 4.0 | 1.3 | 4.8 | 35.7 | 38.1 | 39.1 | 47.4 | 27.8 |
| 15 | 6.3 | 3.6 | 2.5 | 4.1 | 3.5 | 77.6 | 78.4 | 63.9 | 71.5 | 48.1 |
| 16 | 3.0 | 0 | 0 | 0 | 0 | 89.7 | 86.2 | 82.8 | 89.6 | 71.6 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 1.2 | 0 | 0.5 | 1.2 |

**Table 2.** Percentage of deviations from the basic strategy of 100 instances of Model 1 with 10,000 games of training (each) during 1000 further games (each), fitted on Wagenaar's data in Figure 1, with parameters $\alpha = 0.2$, $c = 3$, $att = 0.5$

| | Dealer's Upcard | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Player's total | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 12 | 57.7 | 54.4 | 31.8 | 27.9 | 32.9 | 33.7 | 26.0 | 36.9 | 16.5 | 15.1 |
| 13 | 32.0 | 26.9 | 19.4 | 14.5 | 24.5 | 35.4 | 37.1 | 41.7 | 21.0 | 19.7 |
| 14 | 20.6 | 14.8 | 13.9 | 13.0 | 13.8 | 52.0 | 39.8 | 55.7 | 26.7 | 22.2 |
| 15 | 18.6 | 15.6 | 7.6 | 9.8 | 11.8 | 48.5 | 55.9 | 64.2 | 55.5 | 35.6 |
| 16 | 11.8 | 12.4 | 7.0 | 5.0 | 10.8 | 57.2 | 67.1 | 62.5 | 67.3 | 49.4 |
| 17 | 3.8 | 1.7 | 1.6 | 3.4 | 2.1 | 2.2 | 10.9 | 9.2 | 5.1 | 21.1 |

**Table 3.** Percentage of deviations from the basic strategy of 100 instances of Model 1 with 10,000 games of training (each) during 1000 further games (each), fitted on the basic strategy ($\alpha = 0.1$, $c = 3$, $att = 0.4$)

| | Dealer's Upcard | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Player's total | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 12 | 43.9 | 46.5 | 34.3 | 28.3 | 35.3 | 19.6 | 37.3 | 36.2 | 10.6 | 20.4 |
| 13 | 48.4 | 34.1 | 30.1 | 27.2 | 22.5 | 33.3 | 33.9 | 34.5 | 12.0 | 23.2 |
| 14 | 27.1 | 28.9 | 15.1 | 18.9 | 11.6 | 37.3 | 32.0 | 34.8 | 24.4 | 28.3 |
| 15 | 14.3 | 11.8 | 9.5 | 15.7 | 13.3 | 50.7 | 49.4 | 51.3 | 30.8 | 31.1 |
| 16 | 15.2 | 10.8 | 10.5 | 7.7 | 9.9 | 42.5 | 58.1 | 63.2 | 50.3 | 36.1 |
| 17 | 5.6 | 4.8 | 4.8 | 1.5 | 1.4 | 3.2 | 18.0 | 18.5 | 13.2 | 39.2 |

**Table 4.** Percentage of deviations from the basic strategy of 100 instances of Model 2 with 10,000 games of training (each) during 1000 further games (each), fitted on Wagenaar's data in Figure 1, with parameters $\alpha = 0.2$, $c = 2$, $\gamma = 0.1$

| | Dealer's Upcard | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Player's total | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 12 | 36.8 | 36.5 | 50.0 | 47.2 | 46.7 | 17.8 | 13.9 | 6.8 | 3.0 | 4.4 |
| 13 | 40.7 | 42.0 | 32.8 | 31.8 | 27.4 | 25.5 | 21.6 | 20.2 | 9.7 | 8.6 |
| 14 | 23.7 | 23.0 | 18.5 | 13.2 | 20.9 | 50.4 | 49.1 | 36.2 | 18.7 | 22.3 |
| 15 | 13.8 | 14.7 | 13.8 | 12.6 | 8.7 | 70.0 | 53.3 | 63.8 | 55.0 | 32.4 |
| 16 | 12.5 | 8.2 | 9.7 | 8.2 | 8.1 | 84.8 | 77.4 | 74.3 | 83.3 | 61.9 |
| 17 | 4.9 | 5.0 | 1.6 | 1.0 | 1.8 | 2.1 | 5.8 | 3.1 | 0.7 | 6.5 |

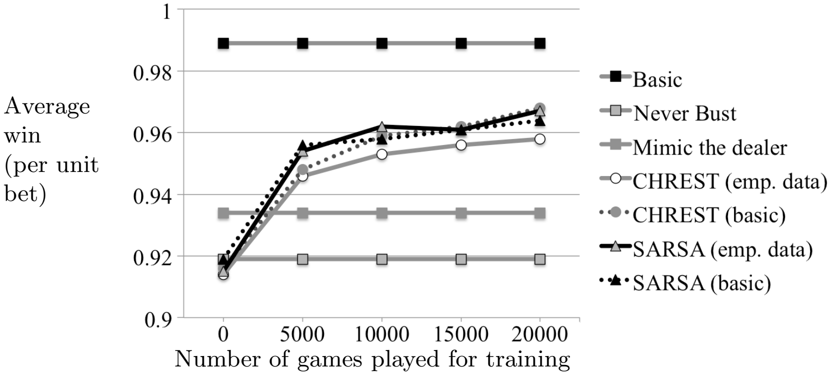

**Fig. 2.** Comparison between average wins of models and fixed strategies, relative to the amount of training

Fig. 2 presents the average wins achieved by the different models (evaluated in batches of 100, on 1000 different games each), relative to the number of games they are trained on, and relative to the unit bet. They perform quite similarly. In particular, their profitability falls in between the simple strategies and the basic strategy – similarly to the casino players in the empirical studies.

## 6 Conclusion

This paper has investigated in how far the behaviour of blackjack players in the casino can be modelled as being the result of learning. Our models were found to generate a behaviour similar to that observed by Wagenaar [4], which is half-way between the basic strategy and a bias to avoid busting. We compared the fit of two different approaches to modelling Wagenaar's data; which are mainly inspired by previous work on conditioning, decision making and SARSA. Whereas the fit of both investigated variants is very encouraging, the fit of the model combining SARSA and softmax turned out to be slightly superior in accounting for Wagenaar's data. However, one needs to be cautious with the interpretation, since our results also highlight the very stochastic nature of the game, which makes model fitting difficult. Furthermore, we used batches of models with the same parameters to model populations of players. In how far the models account for individual differences (e.g. use of explicit strategies) still needs to be tested. This work highlights how difficult it is for a player to learn blackjack by playing (rather than intentionally learning the basic strategy). Our models remain still far from optimal performance as compared to the basic strategy, even with parameters fitted for that goal and 10,000 hands of experience.

An important question raised in this paper is the role of biases on the learning of strategies. We found that a bias towards immediate outcomes (the attribution bias) contributed towards the fit of Model 1 to Wagenaar's empirical data.

The $\gamma$ parameter of Model 2 (the SARSA-variant) has a similar role, since it discounts expected reward relative to immediate reward, and was found to play a crucial role. Future work may address other sources of bias that may play a role in the learning of strategies in casino games such as blackjack. For example, imbalances in the valence of winning a certain amount as compared to losing the same amount, like in the EV model [10], can be incorporated and investigated.

The presented blackjack model mainly hinges on association and reinforcement learning, but not so much on other aspects that CHREST is famous for, such as chunking. Future work may investigate models that rely on both of these kinds of aspects, to model behaviour in games (and other problem solving tasks) with a richer structure of patterns to be memorised and recognised, e.g. poker.

**Acknowledgments.** This work was funded by a grant from the British Academy, under contract number BR100096. We thank three anonymous reviewers for their useful comments.

## References

1. Gobet, F., Lane, P.C.R., Croker, S., Cheng, P.C.H., Jones, G., Oliver, I., Pine, J.M.: Chunking mechanisms in human learning. Trends in Cognitive Sciences 5, 236–243 (2001)
2. Perez-Uribe, A., Sanchez, E.: Blackjack as a test bed for learning strategies in neural networks. In: IEEE International Joint Conference on Neural Networks, IJCNN 1998, vol. 3, pp. 2022–2027 (1998)
3. Kendall, G., Smith, C.: The evolution of blackjack strategies. In: The 2003 Congress on Evolutionary Computation, CEC 2003, vol. 4, pp. 2474–2481 (2003)
4. Wagenaar, W.: Paradoxes of gambling behaviour. Erlbaum, Hillsdale (1988)
5. Thorp, E.: Beat the dealer: A winning strategy for the game of twenty-one: A scientific analysis of the world-wide game known variously as blackjack, twenty-one, vingt-et-un, pontoon or Van John. Blaisdell Pub. Co. (1962)
6. Walker, M., Sturevska, S., Turpie, D.: The quality of play in Australian casinos. In: Finding the Edge: Mathematical Analysis of Casino Games. Institute for the Study of Gambling and Commercial Gaming (2000)
7. Chau, A.W.L., Phillips, J.G., Von Baggo, K.L.: Departures from sensible play in computer blackjack. Journal of General Psychology 127(4), 426–438 (2000)
8. Widrow, B., Gupta, N.K., Maitra, S.: Punish/reward: Learning with a critic in adaptive threshold systems. IEEE Transactions on Systems, Man and Cybernetics 3, 455–465 (1973)
9. Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., Wagenmakers, E.J.: Bayesian parameter estimation in the expectancy valence model of the Iowa gambling task. Journal of Mathematical Psychology 54, 14–27 (2010)
10. Busemeyer, J.R., Stout, J.C.: A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. Psychological Assessment 14, 253–262 (2002)

11. Simon, H.A., Chase, W.G.: Skill in chess: Experiments with chess-playing tasks and computer simulation of skilled performance throw light on some human perceptual and memory processes. American Scientist, 394–403 (1973)
12. Gobet, F., Simon, H.A.: Templates in chess memory: A mechanism for recalling several boards. Cognitive Psychology 31, 1–40 (1996)
13. Ekman, P.: Basic emotions. In: Handbook of Cognition and Emotion. Wiley (1999)
14. Plutchik, R.: Emotion: A psychoevolutionary synthesis. Harper & Row, New York (1980)
15. Rescorla, R.A., Wagner, A.R.: A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A.H., Prokasy, W.F. (eds.) Classical Conditioning II: Current Research and Theory, pp. 64–99. Appleton-Century-Crofts, New York (1972)