In *Proceedings of the 25th International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Cambridge, UK

Acting Irrationally to Improve Performance in Stochastic Worlds

Roman V. Belavkin
School of Computing Science, Middlesex University
London, NW4 4BT, United Kingdom

Abstract

Despite many theories and algorithms for decision–making, after estimating the utility function the choice is usually made by maximising its expected value (the *max EU* principle). This traditional and 'rational' conclusion of the decision–making process is compared in this paper with several 'irrational' techniques that make choice in Monte–Carlo fashion. The comparison is made by evaluating the performance of simple decision–theoretic agents in stochastic environments. It is shown that not only the random choice strategies can achieve performance comparable to the max EU method, but under certain conditions the Monte–Carlo choice methods perform almost two times better than the max EU. The paper concludes by quoting evidence from recent cognitive modelling works as well as the famous decision–making paradoxes.

1 Introduction

During the last several decades, the theory of decision–making under uncertainty has received an extensive treatment by scientists. The most prominent contributions have been made by von Neumann and Morgenstern (1944), Savage (1954), Anscombe and Aumann (1963). Despite the differences in their approach to uncertainty (i.e. objective or subjective), the notion of *utility* has been successfully used to compute the preferences of a decision–maker. Theories such as the dynamic programming by Bellman (1957) and the reinforcement learning partly due to Sutton and Barto (1981) have enabled us to compute the utilities necessary for optimal decision–making. Combined with probabilistic inference (e.g. the Bayes' conditional probability rule), these theories have been used successfully in decision–theoretic agents and robots that can learn autonomously and find solutions to various problems.

Despite these successes, however, soon after its emergence the theory of rational decision–making has been strongly criticised by some psychologists and economists. One simple counter example is the so–called *rational donkey* paradox, when a donkey is placed between two identical haystacks. If the donkey is perfectly rational (i.e. chooses according to the max EU principle), then it will not be able to choose between alternatives with equal EUs. Therefore, some additional mechanism must be involved in choosing, such as a roulette wheel. Moreover, it has been noticed experimentally that human subjects always express some degree of randomness in their choice behaviour even in situations when the choice they make seems irrational according to their previous experience (Myers, Fort, Katz, & Suydam, 1963). The latest cognitive architectures, such as ACT–R (Anderson & Lebiere, 1998), use noise in the utility in order to model the 'imperfect' choice behaviour of humans or animals. Several studies have demonstrated recently that this noisy and 'irrational' component of decision–making may in fact play an important function optimising the behaviour in stochastic environments (Belavkin & Ritter, 2003).

Another famous and powerful counter example to the theory of rational choice has been suggested by Allais (1953) (also known as the Allais paradox) that showed how the theory failed to compute a preference between decisions, which on the other hand was obvious to most of the human subjects. One version of this problem is as follows. Consider a choice between two lotteries:

1. 1/3 chance of winning £300 or 2/3 of not winning anything;

2. A sure win of £100.

One can easily check that two lotteries have equal expected utilities (£100 exactly). Thus, there should be no preference according to the max EU principle. However, most of us (about 70%) would prefer the second lottery demonstrating risk–averse behaviour. Interestingly, when the problem is presented with gains replaced by losses (i.e. loosing money instead of winning), then the preferences of subjects also revert, and a risk–taking behaviour is observed. This paradox has been observed in many experiments using different interpretations. One of the most famous is the study by Tversky and Kahneman (1974), and several theories, such as the *framing* and *prospect* theories (Tversky & Kahneman, 1981), have been proposed to explain these observations. However, most of the theories do not explain the uncertainty that is always present in preferences and choice behaviour of subjects.

In this paper, agents that do not use the max EU principle will be considered. Instead, a Monte–Carlo technique will be used to make decisions randomly (i.e. by drawing samples from probability distributions). These methods will be compared with the more traditional choice strategy by maximising the EUs of decisions. The performance will be analysed using both direct measures of performance as well as information theoretic concepts. Thus, the main focus of this work is the effectiveness of different choice methods, especially in stochastic environments.

A simple decision–theoretic agent architecture and the experimental setup will be described in the first two sections. The results of the tests will be reported in the third section. It will be shown experimentally that although the random methods may lead sometimes to irrational decisions, on average they perform as good as the rational ones, and often significantly outperform them.

The concluding section will discuss the results in the view of psychological evidence as well as the cognitive modelling research. Although resolving the paradoxes of decision–making, mentioned above, was not the goal of this study, some interesting observations will be made that may explain the results observed experimentally.

2 A Simple Decision–Theoretic Architecture

In this section, the design of a very simple decision–theoretic agent will be outlined. This agent will be able to explore its environment, learn and improve its performance according to some criteria. First, let us introduce some notation.

Let $X = \{x_1, \ldots, x_m\}$ be a set of states that an agent can occupy in the world, and let $Z = \{z_1, \ldots, z_n\}$ be a set of actions that the agent can execute. Each action can transfer the agent from one state to another: $x_j = z_k(x_i)$. For convenience of notation, let us denote the set of new states as $Y = \{y_1, \ldots, y_m\}$. Thus, the agent implements a mapping $Z : X \to Y$. In stochastic environments, this mapping is not deterministic and can be described by the probability distribution

$$\boldsymbol{P}(X, Y, Z) = \begin{pmatrix} p^1_{11} & \cdots & p^1_{1m} \\ \vdots & \ddots & \vdots \\ p^1_{m1} & \cdots & p^1_{mm} \end{pmatrix} \cdots \begin{pmatrix} p^n_{11} & \cdots & p^n_{1m} \\ \vdots & \ddots & \vdots \\ p^n_{m1} & \cdots & p^n_{mm} \end{pmatrix} ,$$

where $p^k_{ij} = P(x_i, y_j, z_k)$ is the joint probability of transition from x_i to y_j by executing z_k. In Markov decision problems, matrix (p^k_{ij}) is called the *transition model*. If the agent has no preference between states of the world or actions, then a transition to any state is allowed, and distribution $\boldsymbol{P}(X, Y, Z)$ may be uniform. Let us assume that the agent prefers some states to the others. For example, an agent may loose energy in state i faster than in j, and therefore $i \succ j$ (where $\succ$ denotes binary preference relation). Thus, agent's actions should make transitions to the more preferable states more often.

Traditionally, preferences are expressed by a *utility*, which is a map from states to real numbers $U : X \to \mathbf{R}$. Note, however, that the real numbers are, in fact, not necessary, as only countable sets of states can be ordered by utility. In this paper, we shall consider the uncertainty about utility to be both due to the stochastic nature of the world (i.e. objective uncertainty) and due to the lack of information about its distribution (i.e subjective uncertainty). Thus, we follow the Anscombe and Aumann theory.

Because perception is not in the scope of this paper, let us assume that an agent can ideally recognise the states of the world and which actions it performs. We also assume that the agent can assess correctly the utility of the current state.

The associations between states and actions are recorded by the agent's memory M^k_{ji}, which is a matrix with elements simply counting each transition. The reader should be able to check that after normalisation, the memory M^k_{ji} represents the transition model $\boldsymbol{P}(X, Y, Z)$. Initially, the memory of an agent contains no information. In information theoretic terms, this corresponds to the maximum of entropy $H(X, Y, Z) = E\{-\ln \boldsymbol{P}(X, Y, Z)\}$. The maximum is achieved when $\boldsymbol{P}(X, Y, Z)$ is uniform, and the reader can check that $\max H(X, Y, Z) = \ln(m \times m \times n)$. Note that this information theoretic approach allows us to avoid the argument of objective and subjective probabilities: The prior distribution is defined through the absence of information.

The agent also has a memory for utilities $U(X)$ of the states it has visited. This memory also has no information initially (i.e. utilities of all states are equal). Because no states are preferred, and all transition probabilities are equal initially, the agent starts acting completely randomly. By exploring the world in such a manner, the agent can assess and learn its preferences $U(X)$ (i.e. which states have been 'better' in the past). Consequently, some transitions should become more probable than others, and the entropy $H(X, Y, Z)$ should decrease as a result of changes in probabilities. This change can be evaluated by computing mutual information between variables X, Y and Z:

$$I(X, Y, Z) = H(X) + H(Y) + H(Z) - H(X, Y, Z) ,$$

where $H(X)$, $H(Y)$ and $H(Z)$ are marginal entropies (i.e. for $\boldsymbol{P}(X)$, $\boldsymbol{P}(Y)$ and $\boldsymbol{P}(Z)$), while $H(X, Y, Z)$ is the entropy of joint distribution $\boldsymbol{P}(X, Y, Z)$. If X, Y and Z are statistically independent, then $\boldsymbol{P}(X, Y, Z) = \boldsymbol{P}(X)\boldsymbol{P}(Y)\boldsymbol{P}(Z)$, and $I(X, Y, Z) = 0$. Positive values of $I(X, Y, Z)$ mean that an agent has developed preferences.

Finally, let us consider how the memory of an agent can be optimised in terms of storage requirements. Suppose that there are several states with exactly the same utility: $\exists\, x_1, x_2 \in X : U(x_1) = U(x_2)$. This means that the agent has no preference between these states. We can reduce the size of the transition model M^k_{ji} by considering states only with different utilities: $Y \subseteq X : y_1 \neq y_2 \Rightarrow U(y_1) \neq U(y_2)\ \forall\, y_1, y_2 \in Y$. The cardinality of set Y should be the same as of set U, and it is smaller than cardinality of X. By ordering elements of Y according to U the separate storage for utilities becomes redundant. In this notation, the transitional model implements the Savage approach (i.e. actions map from states to utilities).

In this paper, we shall consider an extreme case when utility divides the world into two subsets of states: S (successes) and F (failures). Thus, $Y = \{S, F\}$. Although this is a crude approximation, it is useful to understand the difference in performance of agents with reduced sets of future states. Note, that such a binary approach has been already successfully used to model human and animal behaviour. The ACT–R cognitive architecture (Anderson & Lebiere, 1998), mentioned earlier, uses the notion of successes and failures to reinforce probabilities of production rules. Many unsupervised and reinforcement learning algorithms also employ binary reward functions.

In the next section, three methods for choosing an action will be presented. These methods are the only architectural difference between the three types of agents tested in this paper.

3 Rational and Irrational Choice

One of the greatest results of probability theory is the relation between conditional probability and joint distribution (known as the Bayes formula)

$$\boldsymbol{P}(X \mid Y, Z) = \alpha \boldsymbol{P}(X, Y, Z) , \tag{1}$$

where α is the normalising constant. As has been mentioned earlier, the associative memory M_{ij}^k of an agent after normalisation represents the joint distribution $\boldsymbol{P}(X, Y, Z)$, which can be used for inference. Indeed, given a transitional model, we can estimate the probability of future outcomes y_j conditional to the current state x_i and actions taken z_k.

3.1 The Maximum Expected Utility

The traditional approach to decision–making is to maximise the expected utility of future states

$$z_k = \arg\max_{z_k \in Z} E\{U\}\ , \text{ where } E\{U\} = \sum_{y_j \in Y} P(y_j \mid x_i, z_k) U(y_j)$$

Agents using this approach behave 'rationally' always choosing what seems to be the best action. The first criticism of this method is that there is no way of choosing an action if expected utilities are equal (and they are at the beginning when no information is available). This problem has been mentioned earlier as the rational donkey paradox. To overcome this limitation, some Gaussian noise of relatively small variance is usually introduced which corrupts the expected utilities by some random values. This approach is used by the ACT–R cognitive architecture (Anderson & Lebiere, 1998), and the noise has allowed for modelling many experiments on human and animals' choice behaviour. The max EU agent, described in the experiments below, used randomness only when the expected utilities were equal. This is equivalent to adding noise of a very small variance.

Another potential drawback we can notice in this method is that only the first moments of utility distributions are used (i.e. the expected values of utilities). The variance and all other potentially useful characteristics of utility distributions are ignored. This may explain the lack of exploration in behaviour of agents using this principle. Indeed, immediately after experiencing the first success, the agent switches to using only the successful action.

3.2 Random Utility

Although small noise can help agents to resolve some problems, it is not clear how large should be the variance of noise corrupting the utilities. Moreover, studies in cognitive modelling of choice behaviour have suggested recently that noise variance should be dynamic. It has been proposed by Belavkin and Ritter (2004) that noise variance should be proportional to the variance of the utility distributions (i.e. the second moments of their distributions). This can be implemented in the following way: Given current state x_i and particular action z_k, the future state y_j can be drawn randomly from its probability distribution, which is inferred using the Bayes formula (1). The utility map U can be used to asses the utility $u_j = U(y_j)$ of such a random state. An action z_k can be chosen by maximising u_j for all actions in Z

$$z_k = \arg\max_{z_k \in Z} u(y_j)\ , \text{ where } y_j \leftarrow \boldsymbol{P}(Y \mid x_i, z_k)$$

This method implements choice by random utilities drawn from their probability distributions, and this allows the agent to act randomly when there is little knowledge about the environment. Indeed, when distributions are close to uniform, their variances are large and there is no clear preference between states. On the contrary, when the agent forms strong preferences, the variance also reduces. Interestingly, this implements a search strategy somewhat similar to the simulated annealing algorithm (Kirkpatrick, Gelatt, & Vecchi, 1983), because the variance is reducing on average, but it also may increase if the agent finds itself in a local maxima.

3.3 Random Action

In this third variation, instead of maximising the utility of future states, the actions are selected directly from their probability distributions. Indeed, we can consider the following conditional probability:

$$\boldsymbol{P}(Z \mid X, Y) = \alpha \boldsymbol{P}(X, Y, Z)$$

Given the utility map U, we can always select future state $y_j \in Y$ maximising the utility (in fact, in our notation Y is a partially ordered set). The action can be drawn from the probability distribution conditional to the current state x_i and the maximum utility state $y_{\max}$

$$z_k \leftarrow \boldsymbol{P}(Z \mid x_i, y_{\max}) \ , \ \text{where } y_{\max} = \arg \max_{y_j \in Y} U(y_j)$$

Again, the agent will choose actions randomly if the joint distribution is uniform. If, however, some actions lead to the maximum utility state more often, then after some training the behaviour should become more 'rational'.

Although there is an obvious difference between the max EU and the random choice strategies, it is not so clear how different are the last two methods. One may notice that cardinalities of sets Y and Z are quite different. Thus, the probability distributions of Y and Z also have different properties and possibly different rates of convergence.

4 Experiment Description

For the purpose of simplicity, the experiments were conducted with environments of small number of discrete states each of which can have a reward (e.g. food) or not. Thus, reward of each state is either 0 or 1. In the experiments, described below, a one–dimensional world of only five states has been used. The agent can also perform only three actions: Stay in the same place, move left or move right.

In this paper, the utility function does not take into account the length of sequence, and therefore we do not consider environmental histories. Again, this is done for simplicity, but the results can be generalised later for utilities that take into account the length of sequence or time. In fact, such a setup is an extreme case of decaying utility with zero decay time.

On each step, the agent records the following information into its memory: The transition m_{ij}^k from state x_i to a new state y_j (or utility) using action z_k. If the agent moves to the state with a reward, then the reward is collected. The rewards can re–appear at different places of the world either randomly or according to some pattern. Three different patterns have been used in the tests:

Random : rewards can occur in any place of the world with equal probability.

Poor : the number of places in the world, where rewards can occur, is smaller than the number of places without rewards.

Rich : the number of places with rewards is larger than the number of places without.

The rates, at which the rewards regenerate in the world, can also be changed. The experiments have been run using several rates of rewards changing from very low to very high rates.

Two main criteria have been used to measure the performance of the agents:

1. The proportion of rewards collected (i.e. a percentage of rewards collected out of all rewards that have appeared).

2. The increase of mutual information between states and actions.

In the next section, the results of tests are reported.

5 Results of the Experiments

Figure 1 compares the performance of three decision theoretic agents in completely random (top), poor (middle) and rich (bottom) worlds. The ordinates on the charts show the percentage of rewards collected by the agents out of all rewards appearing in identical environments and during the same period of time. One can see that very similar performance is achieved by all three agents. For the random pattern (top graph of Figure 1), because the probability of any place containing a reward was the same, all agents have collected similar number of rewards. There seems to be a small variation in the performance when the rate of rewards is the lowest, but this can be explained by the small number of rewards. For all other rates, the performance of all agents is almost identical. For the poor (middle) and the rich (bottom) patterns, the number of rewards collected is greater than for the random world, which indicates that all agents were able to learn where to expect the rewards.

Interestingly, although the number of rewards collected is very similar, the behaviour of agents is very different: The max EU agent most of the time 'preferred' to stay in the same place, and therefore collected only those rewards that appeared in the same place. On the contrary, both random agents have explored the world more and collected the rewards from different places.

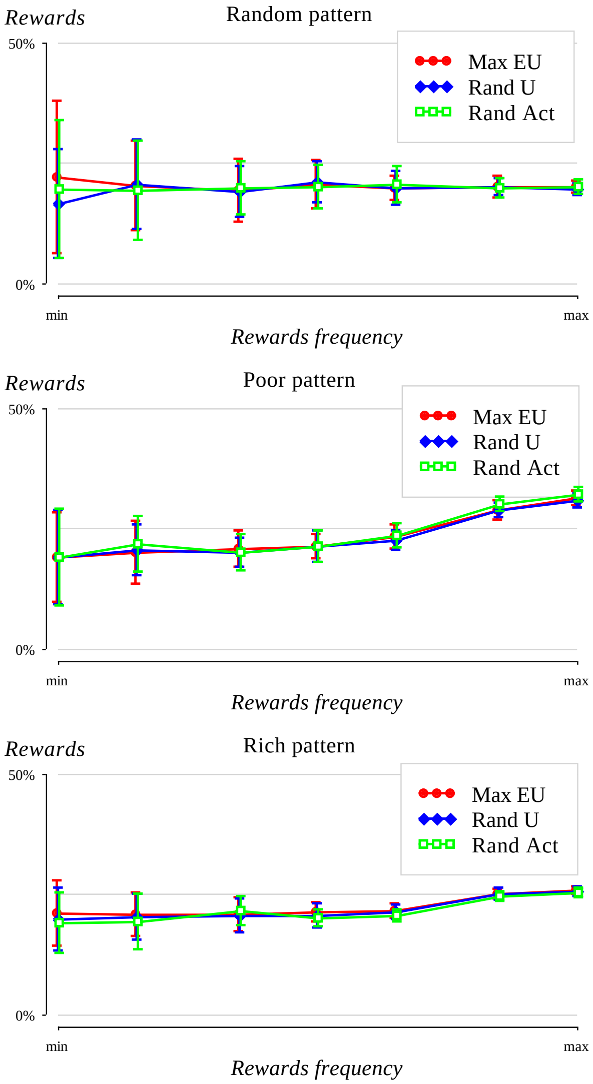

Figure 1: Proportion of rewards collected as a function of rewards frequency for random (top), poor (middle) and rich (bottom) patterns.

One can see that although agents use different tactics for action selection, their performance in terms of number of rewards is very similar. This is an interesting result because two of the agents are not using the traditional max EU principle, and one would expect them to have a disadvantage.

Figure 2 illustrates results of agents with binary representation of future states (i.e. $Y = \{S, F\}$). One can see the dramatic change in performance: Agents collected twice as many rewards as the agents with a full set of future

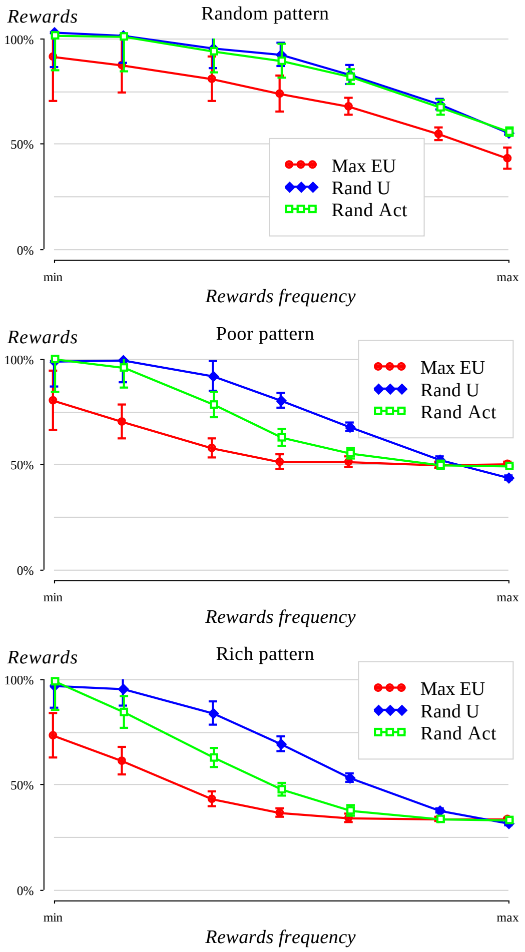

Figure 2: Proportion of rewards collected as a function of rewards frequency for random (top), poor (middle) and rich (bottom) patterns.

states. Perhaps, this result can be explained by the smaller size of the transition model, and hence its ability to learn and re–learn faster.

Furthermore, the charts demonstrate that randomly acting agents have performed better than the max EU agent: At low and medium rates of rewards and rewards occurring according to some regular patterns, the randomly acting agents have collected almost two times more rewards. This result is due to a more explorative behaviour of random agents as opposed to the max EU agent that tends to over–exploit some options. Perhaps, in stochastic worlds with scarce resources, exploration is a more beneficial strategy.

Finally, one may notice that the best performance was demonstrated by the random utility agent. It is not clear exactly why such a result is observed, but

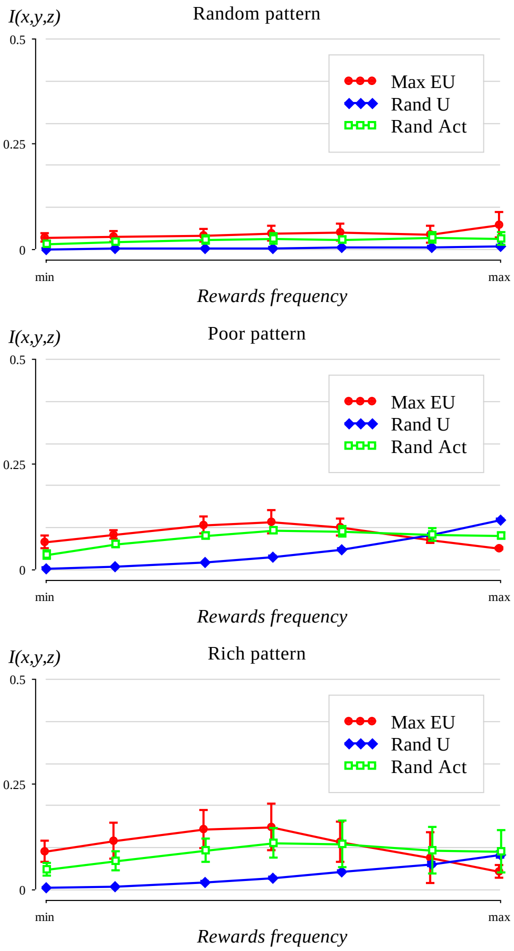

Figure 3: Mutual information acquired as a function of rewards frequency for random (top), poor (middle) and rich (bottom) patterns.

one reason that may be considered is that these agents used three actions (stay, move left and right) and two possible utilities (success and failure). Perhaps, a smaller cardinality of a set contributes to a faster learning of the probability distribution over this set. However, this hypothesis is yet to be tested.

Figure 3 shows the amount of mutual information $I(X,Y,Z)$ accumulated in these tests. One can see that the random utility agent (the one that has the best performance) has acquired the least amount of information in almost all cases. Moreover, the charts show that too much mutual information hinders the performance. Indeed, as was mentioned earlier, positive mutual information reflects the amount of preferences formed by the agent. However, excessive preference to particular states or actions may lead to over–exploitation and not sufficient exploration, which is not a beneficial strategy in stochastic worlds. This is particularly well illustrated by the top graphs of Figures 2 and 3 comparing the performance and mutual information for a completely random world: The max EU agent has the worst performance, but accumulated more information than the 'irrational' agents. This result, however, does not reflect the reality of the world: There is no regular pattern of rewards, and therefore a preference for a particular state is unnecessary.

6 Discussion

In this paper, methods for choosing actions alternative to the traditional maximum expected utility have been tested. The simulations, described here, demonstrated that Monte–Carlo techniques can achieve better performance in stochastic environments because they facilitate a more explorative strategy. Moreover, the balance between exploitation and exploration is maintained due to the characteristics of probability distributions other than the expected values (i.e. moments of higher order than one, such as variance). In addition, the simulations showed that the performance can be improved by reducing the size of the transitional model. This is achieved by considering the set of utilities instead of the set of future states as in traditional Markov decision models.

The idea of dynamic randomness in decision–making proportional to the variance of probability distributions has been discussed recently in the cognitive modelling society: Models that used adaptive dynamic noise in the utilities of production rules matched better the data from studies on animal learning (see Belavkin & Ritter, 2003). The dynamics of noise variance was shown to be proportional to the entropy associated with the success in the task as well as the variance of utilities of the rules. The simulations with such a dynamic control over the uncertainty in decision–making have also achieved better learning and adaptation of behaviour. These results correspond well to the outcomes of the simulations reported in this paper.

An interesting question is whether the random utility theory can also explain why people demonstrate clear preferences in situations when the expected utility theory suggests no preference between decisions, as in the Allais paradox described in the Introduction. Recall that subjects were asked to choose between two alternative lotteries (one with only 1/3 chance of winning £300 and another with a sure win of £100). Although both alternatives had equal expected utilities of £100, the majority of subjects (about 70%) preferred the second lottery. This behaviour can be explained by the random utility choice method, described earlier in this paper. Note that in the first lottery, we should win nothing two out of three times, while in the second lottery we always win £100. Thus, according to the random utility method, two out of three times the random utility of the first option is smaller than that of the second. Remarkably, this proportion also reflects the fact that only about 70% of subjects choose the second lottery, but not all (see Tversky & Kahneman, 1974). The reader may check that the risk–taking behaviour for the reversed version of this problem (i.e. loosing money instead of winning) can also be explained in this fashion.

Although explaining the decision–making paradoxes was not among the main intentions of this research, it is an interesting outcome. It seems that the random decision–making, as opposed to the expected utility theory, promises not only a better performance for agent architectures, but also a better theory for cognitive scientists.

Acknowledgements

I thank both reviewers for very useful comments and suggestions. I also thank Frank Ritter for additional comments and proof–reading of the draft.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'École americaine. *Econometrica*, *21*, 503–546.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Lawrence Erlbaum.

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of Mathematical Statistics*, *34*, 199–205.

Belavkin, R. V., & Ritter, F. E. (2003, April). The use of entropy for analysis and control of cognitive models. In F. Detje, D. Dörner, & H. Schaub (Eds.), *Proceedings of the Fifth International Conference on Cognitive Modelling* (pp. 21–26). Bamberg, Germany: Universitäts–Verlag Bamberg.

Belavkin, R. V., & Ritter, F. E. (2004). Optimist: A new conflict resolution algorithm for ACT–R. In *Proceedings of the Sixth International Conference on Cognitive Modelling* (pp. 40–45). Mahwah, NJ: Lawrence Erlbaum.

Bellman, R. E. (1957). *Dynamic programming.* Princeton, NJ: Princeton University Press.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, J. M. P. (1983, May). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Myers, J. L., Fort, J. G., Katz, L., & Suydam, M. M. (1963). Differential monetary gains and losses and event probability in a two–choice situation. *Journal of Experimental Psychology*, *77*, 453–359.

Neumann, J. von, & Morgenstern, O. (1944). *Theory of games and economic behavior* (first ed.). Princeton, NJ: Princeton University Press.

Savage, L. (1954). *The foundations of statistics.* NY: John Wiley & Sons.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, *88*(2), 135–170.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.