Anticipations Control Behavior: Animal Behavior in an Anticipatory Learning Classifier System

Shorttitle: Anticipations Control Behavior

Martin V. Butz

Department of Cognitive Psychology

University of Würzburg

Würzburg, 97070, Germany

{butz}@psychologie.uni-wuerzburg.de

also

Illinois Genetic Algorithms Laboratory (IlliGAL)

University of Illinois at Urbana-Champaign

Urbana, IL, 61801, USA

{butz}@illigal.ge.uiuc.edu

Joachim Hoffmann

Department of Cognitive Psychology

University of Würzburg

Würzburg, 97070, Germany

{hoffmann}@psychologie.uni-wuerzburg.de

Abstract

The concept of anticipations controlling behavior is introduced. Background is provided about the importance of anticipations from a psychological perspective. Based on the psychological background wrapped in a framework of anticipatory behavioral control, the anticipatory learning classifier system ACS2 is explained. ACS2 learns and generalizes online a predictive environmental model (a model that allows the prediction of future environmental states). The model is a subjective model, that is, no global state information is available to the agent. It is shown that ACS2 can simulate anticipatory learning processes and anticipatory controlled behavior by the means of the model. The simulations of various rat experiments, previously conducted by Colwill and Rescorla, show that the incorporation of anticipations is indeed crucial for simulating the behavior observed in rats. Despite the simplicity of the tasks, we show that the observed behavior reaches beyond the capabilities of model-free reinforcement learning as well as model-based reinforcement learning without online generalization. Possible future impacts of anticipations in adaptive learning systems are outlined.

Keywords

Anticipations

Anticipatory Behavioral Control

Anticipatory Learning Systems

Reinforcement Learning

Predictive Environmental Model

ACS2

1 Introduction

The insight that anticipations influence behavior has been appreciated more and more over the last decades. Cognitive psychology, cognitive sciences, as well as neuro sciences have been discovering anticipatory influences in their diverse research directions. For example, anticipations directly influence the execution of behavior (Kunde, 2001), anticipations influence attentional processes (Pashler, Johnston, & Ruthruff, 2001), or perceptions trigger the anticipation of behavior resulting in behavioral preparedness (Schubotz & von Cramon, 2001).

In artificial intelligence on the other hand, the notion and usage of anticipations for adaptive behavior is still in its infancy. Reinforcement learning (RL) distinguishes now between model-free and model-based reinforcement learning techniques (Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 1998). Model-free RL simulates pure stimulus-response learning and behavior. Model-based RL, on the other hand, builds a predictive model and uses the model to adapt behavior. This approach basically allows planning as well as reacting. The dynamical architecture *Dyna* (Sutton, 1991a) is the most prominent example of this approach. One current major drawback of RL is that usually no online generalization takes place. That is, the RL agent usually does not generalize over the provided perceptions (or sensations, or features) while interacting with the environment. Moreover, the actual benefit or necessity of anticipatory behavior has not been investigated so far.

This work provides evidence for anticipatory behavioral influences in animals and humans, introduces a psychologically motivated behavioral learning theory of anticipatory behavioral control, and analyzes the theoretical constraints in the anticipatory learning classifier system ACS2 (Stolzmann, 1997; Butz, 2002). It is investigated how well ACS2 implements the psychological theory and how the approach allows the realization of anticipatory controlled behavior. It is shown that additional anticipatory learning and behavioral mechanisms can be added easily. Taking the animat approach (Wilson, 1991) to competent adaptive behavior systems, we simulate the behavior of ACS2 in several simple rat experiments. Despite their simplicity, we reveal that the observed rat behavior cannot be simulated with model-free RL since effect associations beyond plain reinforcement values are necessary. Moreover, we reveal that not online generalizing model-based

RL techniques cannot simulate the behavior, either, since modifications in the environment occur during the experiment. We suppose that an online learned and online generalized predictive model representation in combination with anticipatory processes enables strong behavioral competence.

In the next section, we reveal the importance of anticipations and knowledge about anticipatory influenced behavior from a cognitive psychology perspective. Moreover, we provide a framework for anticipatory behavioral control that is consistent with the psychological findings. Section 3 introduces ACS2, providing details to all relevant processes as well as comparing the learning processes to the anticipatory behavioral control framework. Section 4 introduces a first rat experiment and compares behavior of ACS2 with that of the rats. Section 5 studies two further rat experiments. In these experiments additional anticipatory processes are simulated in ACS2 to solve the tasks. Section 6 summarizes and concludes the findings.

2 Anticipations Control Instrumental Behavior: Recent Experimental Evidence in Animals and Humans

A heritage of behaviorism, which restricted itself to objectively observable behavioral phenomena and disregarded any cognitive or even consciousness related explanations of behavior, is that in artificial intelligence learning is mostly considered as being the formation of stimulus-action connections associated with previous reinforcement sensations. This notion is to be traced back to Thorndike's "Law of Effect" according to which the presentation of a reinforcer following an action strengthens a connection between the stimulus or situation present when the action is performed and the action itself so that subsequent presentations of these stimuli elicit the action as a response (Thorndike, 1911).

More recently, though, it became clear that learning theories plainly based on reinforcement are insufficient to explain all observed behavior in cognitive psychology experiments. This section provides evidence for anticipations controlling behavior in animals and humans. Moreover, a framework of anticipatory behavioral control is sketched.

2.1 Evidence in Experiments with Animals

The crucial role of action-outcome relations in instrumental behavior of animals has been already acknowledged by Tolman and his collaborators (Tolman & Honzik, 1930; Tolman, 1932; Tolman, 1949). Tolman's major argument for the insufficiency of traditional behaviorism is the observation of latent learning. In a typical latent learning experiment by Tolman and Honzik (1930), two groups of rats explore a multiple T-maze in several trials with the difference that the first group receives reinforcement (food) at the end of the maze and the second group does not. It is shown, that the rats in the second group move towards the end of the maze faster once food is also provided to them. This shows that the rats must have formed a predictive model representation of their environment which they exploit subsequently to solve an explicit task.

While the diverse latent learning experiments (cf. Thislethwaite, 1951 for an overview) have been subject to several critiques, convincing experimental demonstrations of animal action-outcome learning has been delivered by the use of an outcome devaluation procedure, first employed by Adams and Dickinson (1981). Let us consider an outcome devaluation experiment by Colwill and Rescorla (1985) which we will investigate throughout this work: A group of rats is first trained to perform two different actions which lead to two different outcomes, e.g. lever pressing leads to food pellets and chain pulling leads to a sucrose solution. After training, one of the two outcomes is devalued by associating it with a mild nausea (in this case lithium chloride, LiCl). When the rats are subsequently given the choice between the performance of the two actions in an extinction phase, in which no reinforcement is provided, the animals clearly prefer the action that previously led to the non devalued outcome. Figure 1 shows the experimental setup schematically.

Obviously, the animals did not respond to the situation with any action that was directly reinforced before, but rather expectations of the forthcoming outcome of the available actions led to the avoidance of the previously devalued outcome. Thus, animal behavior is at least partly determined by anticipations of to be expected action outcomes. The result suggests three conclusions: first, the animals have not only acquired stimulus-response ($S-R$) relations, but also some relations about which actions will lead to which outcome, i.e. response-outcome ($R-O$) associations. Second, the acquired $R-O$ representations are involved in

the propagation of the subsequently modified outcome value (devaluation). Third, and most important, the (modified) $R-O$ representations influence the choice of the behavior. In section 5.4 we examine performance of ACS2 in this experiment validating the three suggested conclusions.

In a further experiment, Colwill and Rescorla (1990) examined the impact of the situational context on animal's choice of actions with different outcomes. The assignment of two different outcomes to two different actions were reversed in dependence on the presence of discriminative stimuli (see Figure 2). In one setting, for example, rats received food pellets for pressing a lever and a sucrose solution for pulling a chain in the presence of noise while in the presence of light lever pressing resulted in sucrose and chain pulling resulted in food pellets. After this discrimination training one of the two outcomes, let's say sucrose, was devalued by pairing its consumption with a mild nausea. Finally, the animals were given again the choice between the two actions in the presence of either the noise or the light. They clearly preferred that action under the present stimulus that previously resulted in the non devalued outcome (in our case food pellets). Particularly, the rats preferred lever pressing in the presence of noise whereas they preferred chain pulling in the presence of light. This preference is again unexplainable by stimulus response theories since the devaluation was experienced in the absence of any pressing or pulling action. Simple stimulus associations are not sufficient, either, since the action dependence is the crucial ingredient of the experiment. Thus, as Colwill and Rescorla (1990) argue, the rats have acquired hierarchical $S-(R-O)$ representations which enable them to predict the outcomes of their actions in dependence of the given situation. Consequently, they preferred that action that in the present situation led to the relatively more desirable outcome. Similar behavior in ACS2 is demonstrated in section 5.5.

The impact of $R-O$ relations on animal behavior as well as their conditionalization to discriminative situational contexts have been demonstrated in numerous experiments meanwhile (cf. Roitblat, 1994; Rescorla, 1990; Rescorla, 1991; Rescorla, 1995; Dickinson, 1994; Pearce, 1997). Of course, this does not exclude that also $S-R$ relations as well as $S-O$ relations contribute to the control of animal behavior. However, the available evidence for the acquisition of contingent $R-O$ relations, which are conditionalized to discriminative stimuli if necessary, is by far stronger than the evidence for direct $S-R$ associations. Thus, anticipations of the to be expected outcomes of actions are a central part of animal behavioral control.

2.2 Evidence in Experiments with Humans

The emphasis of the role of action-outcome anticipations in animal action control has its pendant in the classical ideomotor hypotheses (IMH) of human action control. According to the IMH, humans (and animals) select and initiate voluntary actions by an anticipation of their sensory outcomes:

> An anticipatory image, then, of the sensorial consequences of a movement, plus (on certain occasions) the fiat that these consequences shall become actual, is the only psychic state which introspection lets us discern as the forerunner of our voluntary acts (James, 1890, p.501).

Although the IMH was widely acknowledged at the end of the 19th century (Harle, 1861; Lotze, 1852; Münsterberg, 1889), it felt soon into disrepute because the notion that instrumental behavior might be determined by only introspectively available mental states like 'anticipatory images' was not respectable in the upcoming rigorous behaviorism (cf. Greenwald, 1970). Recently however, the IMH experiences a revival in theoretical considerations (e.g. Hoffmann, 1993; Prinz, 1990; Prinz, 1997; Hommel, 1998) as well as in experimental research (e.g. Elsner & Hommel, 2001; Hommel, 1996; Stock & Hoffmann, 2002; Kunde, 2001; Hoffmann, Sebald, & Stöcker, 2001; Ziessler, 1998; Ziessler & Nattkemper, 2001).

For an empirical confirmation of the IMH two things need to be shown: First, when performing goal oriented actions, associations between the performed actions and their contingent sensory outcomes should be formed primarily instead of associations between stimulus conditions and actions. Second, anticipations of the to be expected outcomes should be the forerunners of action initiation. The following exemplar study strongly supports the IMH.

In order to examine the impact of action outcomes as forerunners of action initiation, Kunde (2001) came up with the simple but straightforward idea to explore response-outcome compatibility effects. Participants were required to perform as quickly as possible, for example, a strong key press to a red signal and a soft key press to a green signal. In the compatible response-outcome condition the strong key press was consistently followed by a loud tone and the soft key press by a soft tone. In the incompatible condition the action-outcome assignments were reversed. Although the tones were exclusively delivered *after* the required action

had been initiated, the action-outcome compatibility nevertheless substantially influenced RTs: On average, participants responded about 50ms faster if the required actions resulted in compatible outcomes than if they resulted in incompatible outcomes. Since influences of possible associations between the response signals and the outcome-tones were ruled out by a control experiment,

> [the results] confirm the central assumption of IMH that anticipatory effect representations become endogenously activated for the purpose of response selection (Kunde, 2001, p.393).

To summarize: There is growing evidence that goal oriented behavior in animals as well as in humans is to a great part determined by anticipations of the to be expected outcomes of available actions. Behavioral control by outcome anticipations necessarily presupposes the learning and representation of consistent $R-O$ relations. The integration of discriminative stimulus conditions seems to be a secondary process by which $R-O$ relations become conditionalized, i.e. $S-(R-O)$ representations are formed.

2.3 Anticipatory Behavioral Control: A Tentative Framework

Hoffmann (1993) proposed a tentative framework for the acquisition of behavioral competence that takes the primacy of $R-O$ learning as well as the conditionalization of $R-O$ relations on relevant situational contexts into account. The framework departs from the following basic assumptions (cf. Figure 3):

1. It is supposed that any voluntary action is preceded by an anticipation of to be reached outcomes. Hereby, a voluntary action is defined as performing an action to attain some desired outcome. Thus, a desired outcome, as general and imprecise it might be specified in the first place, has to be represented in some way before a voluntary action can be performed.

2. The real outcome of the act is compared with the anticipated outcome. If there is sufficient coincidence between what was desired and what really happened, representations of the just performed action and the experienced outcomes become interlinked, or an already existing link is strengthened. If there is no sufficient coincidence, no link is formed, or an already existing link is weakened. This formation of integrated action-outcome representations, corresponding to the experienced contingencies, is considered as being the primary learning process in the acquisition of behavioral competence.

3. It is assumed that situational contexts that are present during action performance become integrated into action-outcome representations, if they systematically modify the contingencies between actions and outcomes and/or if a certain action-outcome episode is systematically accompanied by always the same situational context. This conditionalization of sufficiently stable action-outcome relations is considered as being a secondary learning process in the acquisition of behavioral competence.

4. An 'awakening' need or a concrete desire activates the action-outcome representations, in which the outcomes sufficiently coincide with what is needed or desired. Thus, the anticipations of outcomes address the actions that are represented as being appropriate to produce the outcome. If the activated action-outcome representations are conditionalized, the coincidence between the stored conditions and the present situation is checked. In general that action will be performed that in the present situational context most likely produces the anticipated outcome.

5. Conditionalized action-outcome representations can also be addressed by stimuli that correspond to the represented conditions. Thus, a certain situational context in which repeatedly a certain outcome has been produced by a certain action can elicit the readiness to produce this outcome by that action again.

Certainly, the sketched framework is a rather rough one that still requires numerous specifications. However, it integrates many important aspects that are believed to underly the acquisition of behavioral competence into one common framework. First, it takes the well established insight into account that organismic behavior is almost always goal oriented instead of being stimulus driven. Second, as learning is assumed to be driven by comparisons between anticipated and real action outcomes, the anticipations determine which outcomes operate as reinforcers.

Thus, the framework merges classical reinforcement learning, assuming the expectation of a direct reinforcer, as well as latent learning (or learning of a predictive environmental model), assuming the expectation of outcomes without immediate value. Third, the framework considers the recent evidence that anticipations of outcomes are indeed forerunners of action initiation, as has been already proposed more than hundred years ago. Finally, also stimulus driven habitual behavior is covered, as it is assumed that representations of action-outcome relations become conditionalized to the typical contexts in which they are experienced.

As can be seen, the framework comprises many different aspects of cognitive psychology research and consequently appears to be a useful departure point for further experimental and simulation studies of organismic behavioral learning.

3 ACS2

The anticipatory learning classifier system ACS (Stolzmann, 1997) was originally intended to simulate and evaluate Hoffmann's learning theory of anticipatory behavioral control (Hoffmann, 1993). ACS is a rule learning system that learns and generalizes online a predictive model of its environment. Each rule, or *classifier*, consists of three basic parts, a condition, an action, and an effect. The complete set of those classifiers, the *population*, represents the complete current knowledge about the environment. Reinforcement learning techniques are applied to adapt behavior.

This section introduces ACS2, the current state-of-the-art of ACS including genetic generalization and further modifications. Moreover, ACS2 is compared to the theory of anticipatory behavioral control. First, background of related artificial learning systems is provided.

3.1 Background

All learning systems that represent and utilize predictions of future states to adapt behavior are related to ACS2. One of the first approaches in this respect was pursued in Sutton's dynamical architecture *Dyna* (Sutton, 1991b). In Dyna an environmental model is learned for the further improvement of RL capabilities. With the learned environmental model, also anticipatory behavioral processes can be simulated. While previous Dyna approaches usually explicitly store each experienced situation-action-resulting situation triple with statistics, ACS2 generalizes online over perceptual attributes. Thus, ACS2 is basically the next step in the general Dyna architecture. Several algorithms and processes introduced in this work actually stem from work on Dyna.

Holland (1990) proposed a somewhat similar idea in the learning classifier system framework (Holland, 1976; Lanzi, Stolzmann, & Wilson, 2000). The idea is to include tags in the message list (comparable to a feature vector) that allow the distinction of predictions, perceptions, actions, and so forth. Riolo

(1991) integrated this concept in his CFSC2 system showing that the system is able to form a predictive environmental model and use the model to adapt behavior. However, CFSC2 did not apply any generalization mechanisms so that the learning classifier system spirit was somewhat lost. Moreover, the tags appear hard to handle and seem to cause more interference than benefit.

Other related systems with predictive environmental model representations include model learning artificial neural networks (NNs) as well as anticipatory learning classifier systems (ALCSs). On the NN side, for example, Tani (1996) succeeded in the simulation of model-based learning on a mobile robot platform. His recurrent neural net (RNN) succeeded in diminishing the state prediction error. Moreover it was shown that planning was possible once the model was present in the RNN and the net was situated in the environmental context. Problems appeared to be scalability and reliability of the model learning approach as well as the difficulty of determining the accuracy of the predictions.

On the ALCS side, Drescher (1991) provided a first approach (not yet terming the system an ALCS). Based on the Piagetian development theory he developed a *schema mechanism* that forms a generalized environmental model online. He was able to show interesting developmental stages in his system drawing relations to the Piagetian theory of development. However, his system did not prove to be robust as can be observed by his limited experimental results. The further investigation of Drescher's ideas, however, seems worthwhile.

Recently, another ALCS termed YACS has been introduced which applies different learning mechanisms but evolves a similar model (Gérard & Sigaud, 2001b). Also a generalization mechanism was added to YACS which proved to evolve maximally compact environmental representations (Gérard & Sigaud, 2001a). It is necessary to further study the differences between YACS and the current ACS2 system. While more publications have been made with ACS2 and more problems solved, YACS showed to solve certain maze tasks with a smaller number of overall classifiers. The size of the environmental model, however, is similar in both systems.

Another ALCS system is the *dynamic expectancy model* (Witkowski, 1997). The system builds an environmental model consisting of rules similar to ACS2. However, although Witkowski mentions a general-

ization mechanism, the mechanism has not been applied in the provided results. Interesting animat behavior has nevertheless been shown as, for example, extinction behavior in Witkowski (2000).

Finally, we want to mention Robert Rosen's contribution to the notion of anticipations. His book anticipatory systems (Rosen, 1985) was the first contribution that approached anticipations from a mathematical perspective. Later, Rosen sees anticipations as a necessary ingredient in the manifestation of life (Rosen, 1991). Anticipations allow a new kind of complexity which is mandatory for living beings. These propositions might sound rather strong and we do not pursue them any further herein. We rather intend to contribute to the general idea and importance of anticipations in adaptive behavior. For this, we give an overview over the investigated ACS2 system in the next sections.

3.2 Agent and Knowledge Representation

Similar to other agent architectures, ACS2 autonomously interacts with an environment. In a *behavioral act* at a certain time t, the agent perceives a current situation in the form of sensory attributes (or a feature vector) $\sigma(t) \in \{\iota_1, ..., \iota_m\}^L$, the agent then acts upon the environment with one of the predefined actions $\alpha(t)$, the environment provides scalar reinforcement $\rho(t)$, and the next behavioral act begins.

While interacting, ACS2 iteratively learns a predictive model of the encountered environment. The model is represented by a population $[P]$ of condition-action-effect rules, i.e. the classifiers. Each classifier predicts action effects given the specified condition. A classifier in ACS2 always specifies the state of all resulting sensory attributes. It consists of the following main components.

- *Condition part* (C) specifies the set of situations in which the classifier is applicable.
- *Action part* (A) proposes a possible action.
- *Effect part* (E) predicts the effects of the proposed action in the specified conditions.
- *Quality* (q) measures the accuracy of the predicted effects.
- *Reward prediction* (r) estimates the long-term reinforcement encountered after the execution of action A in condition C.

- *Immediate reward prediction* (ir) estimates the direct reinforcement encountered after execution of action A in condition C.

The condition and effect part consist of the values perceived from the environment and '#'-symbols (i.e. $C, E \in \{\iota_1, ..., \iota_m, \#\}^L$). A '#'-symbol in the condition, called *don't-care symbol*, denotes that the classifier matches any value in this attribute. A '#'-symbol in the effect part, called *pass-through symbol*, specifies that the classifier predicts that the value of this attribute will not change after the execution of the specified action. Non pass-through symbols in E anticipate the change of the particular attribute to the specified value. The action part specifies any action possible in the environment. The measures q, r, and ir are scalar values where $q \in [0, 1]$, $r \in \Re$, and $ir \in \Re$. A classifier with a quality q greater than the reliability threshold θ_r (usually set to 0.9) is called *reliable* and becomes part of the internal environmental model. A classifier with a quality q lower than the inadequacy threshold θ_i (usually set to 0.1) is considered as inadequate and is consequently deleted. The immediate reward prediction ir is separated from the usual reward prediction r in order to enable proper internal reinforcement learning updates. All parts are modified according to a reinforcement learning mechanism, and according to two model learning mechanisms specified in section 3.3.

Additionally, each classifier comprises a *Mark* (M) that records the values of each attribute of all situations in which the classifier did not predict correctly sometimes. The *mark* has the structure $M = (m_1, ..., m_L)$. Each attribute $m_i \subseteq \{\iota_1, ..., \iota_m\}$ records all values at position i of perceptual strings in which the specified effect did not take place after execution of action A. Moreover, each classifier specifies a *GA time stamp* t_{ga}, an *ALP time stamp* t_{alp}, an *application average* aav, an *experience counter* exp, and a *numerosity* num. The two time stamps record the time of the last learning module applications. The application average estimates the frequency with which the classifier is updated (i.e. part of an action set). The experience counter counts the number of applications. The numerosity denotes how many identical classifier this *macroclassifier* represents.

3.3 Learning Processes

Initially, classifiers are mainly generated by a covering mechanism in the anticipatory learning process (ALP). Later, the ALP generates specialized offspring from over-general classifiers while a genetic generalization

process produces generalized offspring. RL techniques are applied to the evolving rules forming a behavioral policy in the evolving environmental model.

Figure 4 illustrates the interaction of ACS2 with its environment and its learning application in further detail. After the perception of the current situation $\sigma(t)$, ACS2 forms a match set $[M]$ comprising all classifiers in the population $[P]$ whose conditions are satisfied in $\sigma(t)$. Thus, $[M]$ holds the complete predictive knowledge for the current situation. Next, an action $\alpha(t)$ is chosen according to the applied behavioral policy. Herein, a simple $\epsilon-$greedy strategy is applied as often used in RL (Sutton & Barto, 1998). With respect to $\alpha(t)$, an action set $[A]$ is generated that consists of all classifiers in $[M]$ whose action equals $\alpha(t)$. Thus, $[A]$ comprises the predictive knowledge restricted to the chosen action given the current situation. After the execution of $\alpha(t)$ and the reception of reinforcement $\rho(t)$, classifier parameters are updated by the ALP and the applied RL technique and new classifiers might be generated as well as old classifiers might be deleted by the ALP and the genetic generalization process.

The basic learning mechanisms are two interacting model learning mechanisms as well as one reinforcement learning mechanism. The anticipatory learning process (ALP) is the specializing component of the model learning mechanism. The ALP evaluates rules and detects which rules are over-general. Once an over-general rule is detected, specialized offspring is generated. Genetic generalization, on the other hand, is an indirect generalization procedure. Accurate classifiers are chosen for generating generalized offspring. In turn, over-specialized as well as inaccurate classifiers are deleted.

3.3.1 Anticipatory Learning Process

The ALP updates the quality q, the mark M, the ALP time stamp t_{alp}, the application average aav, and the experience counter exp. The quality q is updated according to the classifier's anticipation. If a classifier correctly specified changes and non-changes, called *expected case*, its quality is increased ($q \leftarrow q + \beta(1-q)$). If the classifier specifies an incorrect effect, termed *unexpected case*, its quality is decreased ($q \leftarrow q - \beta q$). Parameter $\beta \in [0,1]$ denotes the *learning rate* of ACS2.

Additional to the parameter updates, the ALP generates specialized offspring and/or deletes *inaccurate* classifiers. Specialized classifiers are generated in two cases. In the *expected case*, a classifier might be

generated if the mark M differs from the situation $\sigma(t)$ in some attributes. This means that the classifier previously encountered situation(s) (characterized by the mark) in which its predictions were incorrect. Thus, the condition of the new classifier is specialized in those differing attributes. In the *unexpected case*, a classifier is generated if the effect part of the classifier can be further specialized (by changing pass-through symbols to specific values) to specify the perceived effect correctly. All positions in condition and effect part are specialized that change from $\sigma(t)$ to $\sigma(t+1)$.

A classifier is also generated if there was no classifier in the actual action set $[A]$ that anticipated the effect correctly. In this case, *covering* applies in which a classifier is generated that is specialized in all attributes in condition and effect part that changed from $\sigma(t)$ to $\sigma(t+1)$.

The attributes of the Mark M of a new classifier are initially empty. Quality q is set to 0.5 in the covering case and is inherited from the parental classifier (minimally set to 0.5) in the other reproduction cases. Reward prediction r and immediate reward prediction ir are set to 0 in the covering case but are inherited from the parent in the other cases. For further details on the learning process please refer to (Stolzmann, 2000; Butz, 2002).

3.3.2 Genetic Generalization Mechanism

While the ALP specializes classifiers in a quite competent way, over-specializations can occur sometimes as studied in (Butz, 2002). Since the over-specialization cases can be caused by various circumstances, a genetic generalization (GG) mechanism was applied that, interacting with the ALP, results in the evolution of a complete, accurate, and maximally general model. The basic framework of the genetic algorithm was derived from Wilson's accuracy based learning classifier system XCS (Wilson, 1995). The mechanism works as follows.

After the application of the ALP, it is determined if the mechanism should be applied. Classifiers are reproduced in the action set $[A]$ proportionally to their quality value q. Reproduced classifiers are crossed and mutated in the conditions. Hereby, a generalizing mutation is applied that randomly changes specialized attributes back to *don't-care symbols*. If a generated classifier already exists in the population, the new classifier is discarded and if the existing classifier is not marked its numerosity is increased by one. If no

identical classifier exists, the quality q of the new classifier is decreased by 0.5 and it is inserted in the population. If an action set $[A]$ exceeds the action set size threshold θ_{as}, excess classifiers are deleted in $[A]$. Deletion causes the extinction of low-quality as well as over-specialized classifiers.

3.3.3 Subsumption

To further emphasize a proper model convergence, subsumption is applied similar to the subsumption method in XCS (Wilson, 1998). If a new classifier is generated, regardless if by ALP or GG, the set is searched for a subsuming classifier. The new classifier is subsumed if a classifier exists that is more general in the conditions, specifies the same effect, is *reliable* (its quality is higher than the threshold θ_r), is not marked, and is *experienced* (its experience counter exp is higher than the threshold θ_{exp}). If there exists more than one possible subsumer, the subsumer with the most *don't-care symbols* is chosen. In the case of a draw, the subsumer is chosen at random. If a subsumer was found, the new classifier is discarded and either quality or numerosity of the subsumer is increased dependent on if the new classifier was generated by ALP or GG, respectively.

3.3.4 Interaction of ALP and GG

Several distinct studies in various environments revealed that the interaction of ALP and GG is able to evolve a complete, accurate, and maximally general model in various environments in a competent way (cf. Butz, Goldberg, & Stolzmann, 2000; Butz, 2002). The basic idea behind the interacting model learning processes is that the specialization process extracts as much information as possible from the encountered environment continuously specializing over-general classifiers. The GG mechanism, on the other hand, randomly generalizes exploiting the power of a genetic algorithm where no more additional information is available from the environment. The ALP ensures diversity and prevents the loss of information of a particular niche in the environment. Only GG generates identical classifiers and causes convergence in the population.

3.4 Behavioral Policy

The behavioral policy of ACS2 is directly represented in the evolving model. Each classifier specifies the reward prediction estimate r and the immediate reward prediction estimate ir which control behavior. Thus, the reward estimates are dependent on the structure of the classifiers so that the environmental model as a whole needs to be specific enough to prevent misleading averaging of the estimates. Only if no averaging takes place it is assured that the classifier population can represent an optimal policy within the predictive model. If averaging takes place, *model aliasing* (Butz, 2002) might prevent the evolution of an optimal policy as previously identified in different contexts (e.g. Whitehead & Ballard, 1991; Dorigo & Colombetti, 1997).

As visualized in Figure 4, the reward related parameters r and ir are updated after the action was executed, the next environmental situation was perceived and the subsequent match set was formed. The update combines immediate reinforcement $\rho(t)$ with discounted future reward.

$$r \leftarrow r + \beta(\rho(t) + \gamma \max_{cl \in [M](t+1) \wedge cl.E \neq \{\#\}^L} (cl.q \cdot cl.r) - r) \tag{1}$$

$$ir \leftarrow ir + \beta(\rho(t) - ir) \tag{2}$$

Again, parameter $\beta \in [0, 1]$ denotes the learning rate biasing the estimates more or less towards recently encountered reward. Parameter $\gamma \in [0, 1)$ denotes the discount factor similar to Q-learning (Watkins, 1989). In contrast to Q-learning, however, the rules may be applicable in distinct situations. The values of r and ir consequently specify an average of the resulting reward after the execution of action A over all possible situations of the environment in which the classifier is applicable. Thus, model aliasing can take place.

Usually, ACS2 applies a simple ϵ-greedy action selection strategy. An action is chosen at random with a probability ϵ and otherwise the best action is chosen. The action specified by the classifier with the highest qr value in a match set $[M]$ is considered the best action in ACS2.

The behavioral capabilities of ACS2 will be of major interest in the following sections. We will show how the reinforcement values can be propagated as well as how policy execution can be modified to generate anticipatory behavior. The next section, however, first compares ACS2 to the theory of anticipatory

behavioral control.

3.5 Relation to Anticipatory Behavioral Control

With a picture of ACS2 in hand, we can now compare the learning and policy mechanisms in ACS2 to the theory of anticipatory behavioral control, introduced in section 2.3. All five theoretical points are addressed.

The first point addresses the representation of predictions before action execution and their control of behavior. Before action execution, ACS2 always forms a match set $[M]$ which represents all predictive and reinforcement knowledge in the current situation. Explicit anticipations can be formed using the information in $[M]$. Thus, an action can be selected that promises to lead to a desired outcome. Section 5.3 introduces a mechanism that explicitly generates predictions beforehand and causes explicit anticipatory behavior as stated in the theory.

The second point matches perfectly with the ALP. In the ALP, anticipated outcomes are compared to real outcomes. Moreover, action-effect relations become interlinked or strengthened by generating a new classifier by the covering mechanism or by increasing the quality of an old classifier. In ACS2, one coincidence is enough to form the link. Other more constrained methods seem possible. Insufficient coincidences are weakened as manifested in the quality decrease and deletion of classifiers.

The third point addresses the consideration of situational context which is, according to the theory, integrated into the action-effect relations. The consideration of situational context is realized in the marking process and the consequent specialized offspring generation in an expected case. Since the specialization of action-effects always comes first, this further conditional specialization is indeed a secondary process in ACS2.

The fourth point is not realized in its explicit form in ACS2. ACS2's action policy is still mainly stimulus-response driven since it is mainly based on the reinforcement prediction r. However, also the predictive knowledge has an influence as manifested in the additional consideration of the quality q in the determination of the current best action as well. Section 5 shows how anticipations can directly influence action selection or anticipations can influence action selection mediated by an alternation in the reward

prediction value r.

The fifth point, is the predominant factor determining behavior in ACS2. Stimuli trigger match set generations which trigger action selection. Thus, conditionalized action-outcome relations are indeed addressed by stimuli.

Generalization is not explicitly addressed in the learning framework. In ACS2, genetic generalization can be compared to a continuous weakening of conditions in action-effect bonds. Although generalization has not been considered in the theory, the mechanism contributes strongly to model condensation and generalization. It can be regarded as a general process of forgetting unimportant details. Subsumption has a similar effect. Reinforcement learning is still a relict of the previous behavioristic thinking and might eventually become completely modified in ACS2. In this work, though, we show that anticipatory controlled behavior can already be simulated in the current framework.

Although certainly not a perfect match, many points of the proposed framework are implemented in ACS2. Section 5 introduces additional mechanisms that make the system match even closer. We want to note that ACS2 is certainly not the only possibility to satisfy the framework nor might it be the best one. However, it is a framework that matches closely. Further advantages of ACS2 are the great generalization capabilities as well as the rule-based representation which allows explicit knowledge extraction. As will be seen below, generalization and rule representation enable or at least facilitate further research on the idea of anticipatory behavior. The next sections show how animal behavior matches with behavior simulated with ACS2 as well as how anticipatory cognitive processes can control behavior in ACS2.

4 Stimulus Dependent Response-Effect Relations

In order to validate the idea of anticipations controlling behavior, ACS2 is now and in the next section compared to results of three psychological experiments previously conducted with rats. The intention is to show that ACS2 is able to mimic animal behavior as well as that anticipatory behavioral control is necessary for simulating similar behavior. Performance of ACS2 is compared to other artificial learning frameworks as well.

This section introduces a first rat experiment published in Rescorla (1990). The experiment is simulated and performance of ACS2 is evaluated.

4.1 The Rat Experiment: Hierarchical $S-(R-O)$ Relations

The major intention of Rescorla (1990) was to evaluate whether or not hierarchical stimulus-(response-effect) ($S-(R-O)$) relations are formed in rats. In order to evaluate this suspicion, Rescorla trained rats with a standard procedure teaching various ($R-O$) relations with respect to discriminative stimuli.

Figure 5 visualizes the experimental setup. The experiment was subdivided into three stages. During the first stage, each animal was trained with three stimuli (i.e. light=L, noise=N, and tone=T) in which two different responses (i.e. pressing a lever or pulling a chain) were reinforced with one of two outcomes. In the presence of light (L), the associative $R-O$ relations R_1-O_1 and R_2-O_2 were in effect each of which were also in effect in one of the auditory stimuli. Thus, during the first stage, L shared the R_1-O_1 relation with N and it shared the R_2-O_2 relation with T. Furthermore, inter trial intervals (ITI) were presented in which no stimulus was present and no action had any effect. During stage two, neither action had any effect and only light stimuli were presented. Thus, the learned R_1-O_1 and R_2-O_2 were extinct during that stage. In the first and second stage, always either chain or lever were present but not both. Finally, in stage three the actions of the rats were monitored under the auditory stimuli in the presence of chain and lever. Actions again did not cause any effect. Rescorla (1990) supposed that only if the rats form hierarchical $S-(R-O)$ relations, the extinction phase could affect the preference during the test phase to execute that action that previously produced a not extinct $R-O$ relation. For further details on the rat experiment the interested reader is referred to the cited article.

The suspicion was confirmed. Rats significantly prefer that action that resulted in the $R-O$ relation during phase one that was not extinct during phase two as depicted in Figure 6. Thus, $R-O$ relations must have been formed which are extinct rather independent of situational context. It was also observed that the mean response time during stimulus presentation is significantly higher than during the ITI which confirms that the rats learned that a stimulus is necessary to successfully obtain reinforcement. Moreover, during the second half of the test phase the response frequency declined significantly on stimulus presentation.

4.2 Simulating the Experiment with ACS2

To specify the simulation of the rat experiment with ACS2, we need to define perceptions, actions, reinforcement, and the length of each experimental stage. We code the experiment as a two-step environment in which ACS2 can act upon one of the manipulanda (lever or chain) in the first step and then consume the resulting reinforcer in the second step. Each time a reinforcer was consumed, the environment is reset and a new stimulus is presented.

Situations are coded as a binary string of length seven denoting the presence of food, sucrose, lever, chain, noise, light, and tone. ACS2 can execute four actions: "pull", "push", "eat", and "do nothing". A reinforcement of 1000 is provided when eating food pellets or sucrose. The length of the three phases are determined according to the steps that are experienced by the rats. During the training phase, the rats were trained for 20 successive days with two training sessions per day with 16 trials with light and 8 trials each with noise and with tone per session. These are all in all 1280 situation action result combinations. The same number of combinations was presented to ACS2. In one half of the combinations, light was present and in the other half, either noise or tone was present. During the subsequent six days, the rats received two extinction sessions per day in which 16 light presentations took place. Thus, 192 extinction trials were presented to ACS2. Finally, during the test session the rats received four presentations each of the two auditory stimuli. In this test phase, the mean response per minute was monitored. For ACS2 we presented 160 cases and monitored behavior for each. For now, inter trial intervals (ITIs) were not presented to ACS2. The effect of ITIs on ACS2 in this experiment is discussed later.

It seems obvious that ACS2 will not be able to exhibit behavior comparable to the behavior of the rats. As noted above, ACS2's behavior is basically stimulus-response driven. The reinforcement extinction causes a decrease in the reward prediction values in the respective classifiers. Actions are selected according to the reward prediction of classifiers. Effects are only taken into account implicitly. Thus, how can the supposedly hierarchical $S-(R-O)$ relation be devalued?

However, as Figure 6 reveals, ACS2 is able to learn the $S-(R-O)$ relation. It exhibits behavior quite

similar to the rats. [1] ACS2 makes the distinction and preferentially executes the action that is part of the not extinct $R-O$ relation. Also, the difference between different and same $R-O$ combination diminishes later in the test phase. According to Rescorla, the differences in the rat behavior did not reach significance anymore. Moreover, the decrease in performance frequency can be observed in ACS2: the "do nothing" and "eating" actions are executed increasingly often during the test phase.

The differentiating behavior emerges from the generalization process and the consequent generalized environmental representation in combination with the reinforcement representation in the classifiers. ACS2 generates classifiers that specify accurate action-effect relations with maximally general conditions. In this experiment, ACS2 forms a classifier that specifies that if either L or N is present, O_1 will follow R_1. Similarly, it forms a classifier that specifies that if either L or T is present, O_2 will follow R_2. When R_1 and R_2 are now devalued in the L condition, R_1 is consequently also devalued in the N condition and R_2 is devalued in the T condition. Thus, ACS2 makes the distinction.

For a classifier to represent L *or* N in its condition part in the chosen coding, it can only specify $\neg T$ since an explicit *or* representation is not possible in the conditions of classifiers in ACS2 right now. Thus, the result is only obtainable if no ITI is simulated. In a simulation with ITI, $\neg T$ is also applicable in the ITI and consequently not sufficient to represent the relation. This suspicion was confirmed in experiments with ITI in which ACS2 does not exhibit any differentiation between the same and different $R-O$ relations. Moreover, when not applying genetic generalization in the setting without ITI, the result was not achievable, either. The anticipatory learning process usually generates the individual classifiers as well as the classifier with condition $\neg T$. In the test phase, the classifier that specifies $N-R_1-O_1$ overrules the more general but devalued classifier $\neg T-R_1-O_1$ so that the distinction does not apply.

Several important observation were made in this simulation. First, ACS2 exhibits an implicit $S-(R-O)$ structure since it differentiates between same and different $R-O$ relations in dependence of S in the test phase. Second, emergent behavior results from the interaction of the reinforcement representation in classifiers and the online generalized model. Although the generalized representation might not be comparable

[1] The parameters in ACS2 were set to: $\beta = 0.05$, $u_{max} = \infty$, $\gamma = 0.95$, $\theta_{ga} = 10$, $\mu = 0.3$, $\chi = 0.8$, $\theta_{as} = 20$, $\theta_{exp} = 20$, $\epsilon = 0.4$. The ACS2 results are averaged over 1000 experiments. Similar results were obtained with variations in θ_{ga} and ϵ, the two most influential parameters in ACS2.

to the rats (the rats most probably did not specify that if not T then $R_1 - O_1$ but rather if L or N, then $R_1 - O_1$), it showed that the $S - (R - O)$ structures can also be obtained without any explicit hierarchical structure. Finally, the results were obtained independent of parameter settings. Thus, the results point to the plausibility of the learning mechanism and the theory of anticipatory behavioral control.

As a final point it is interesting to see how other learning systems would behave. In model-free RL approaches as well as model-based RL approaches without online generalization the transfer would not be possible at all since training, extinction, and test phase differed in the setup structure (either lever or chain was present during training and extinction but both were present during testing). However, even if the simulation would have been conducted in a way that always both manipulanda were present, model-based RL would not be able to show similar behavior since it would learn all situation-action-effect relations exemplarily. For online generalizing model-free RL mechanisms such as previous learning classifier systems (Holland, 1976; Lanzi, Stolzmann, & Wilson, 2000), the system would not distinguish between the different outcomes and backpropagate simple reinforcement. Thus, a learning classifier system would not distinguish between the outcomes. The comparison stresses the importance of a predictive model representation in combination with online generalization. Moreover, it points out the necessary distinction between conditions, actions, and effects. Only due to the conditionalized generation of action effect associations could behavior match with the rat behavior.

5 Explicit Anticipations Influence Behavior

While the previous section showed emergent anticipatory behavior in ACS2, this section shows how the evolving generalized environmental model can be used to distribute reinforcement internally. It is shown that reinforcement values can be adapted to draw conclusions that are appropriate but would have not been possible without the generalized anticipatory model. In more psychological terms, it is shown that ACS2 is able to use its internal generalized environmental model for distinct cognitive processes that allow a "mental" adaptation of behavior.

The study herein is mainly based on the work published in Stolzmann, Butz, Hoffmann, and Goldberg (2000). Due to the changes from ACS to ACS2, though, some parts of the additional mechanisms have

changed. Moreover, genetic generalization is applied throughout. To evaluate the mental adaptation possibilities, ACS2 is tested in a simulation of the two rat experiments published by Colwill and Rescorla (1985) and Colwill and Rescorla (1990) introduced in section 2.1.

This section recapitulates the response-effect experiment by Colwill and Rescorla (1985) and stresses its peculiarity. Next, anticipatory mechanisms are introduced to ACS2 to enable the system to draw mental conclusions. Finally, performance of ACS2 is revealed in the simulation of Colwill and Rescorla (1985) as well as in the simulation of the harder stimulus-response-effect experiment (Colwill & Rescorla, 1990).

5.1 Response-Effect Learning Task

The herein investigated response-effect learning task was originally done with rats by Colwill and Rescorla (1985). Section 2.1 already revealed the basic implications of the experiment. The intention was to investigate if and in what way rats evolve response-effect ($R - O$) relations.

Figure 1 gives an abstract view of the experiment. Rats were tested in a three stage experiment. First, they were taught to execute two distinct possible actions R_1 and R_2 (pressing a lever and pulling a chain). One action led to one type of (positive) reinforcer (sucrose) and the other one to the other (positive) reinforcer (food pellet). Next, without the presence of lever or chain, reinforcers were provided separately and one of the reinforcers was devalued. Finally, the rats were tested if they would choose to press the lever or pull the chain, which were simultaneously present during testing. All three slightly different experimental settings in the original work showed that the rats were preferring the action that previously led to the non-devalued reinforcer during the test phase. Figure 7 shows the performance of the rats during the test phase in all three settings. Additional to the observed successful distinction during testing, the rats also showed to decrease response frequency during testing. Moreover, sucrose was always more appealing than food pellets. Finally, also in the last experiment, in which one reinforcer was sated, the rats showed the basic distinction. Only motivational influences, that is, the motivation to go for the not sated reinforcer, could have triggered the difference in this case.

The experiment shows that rats must have formed context independent response outcome associations that control behavior. Once an outcome is devalued, the associations that lead to the devalued outcome are

(possibly implicitly) devalued as well so that the rats prefer to execute that action that led in phase one to the outcome that was not devalued in phase two.

This outcome dependent action selection can be obtained neither by any model-free RL mechanism, nor by model-based RL approaches without online generalization. Model-free RL fails since it relies on a direct interaction with the environment for learning but the connection "action (pressing or pulling) leads to the devalued reinforcer" is never encountered online. Model-based RL can learn this association since reinforcement can be propagated internally by the means of the learned predictive model (e.g. Sutton, 1991b). However, only model-based approaches that generalize online over perceptual attributes are able to solve the transfer task since each experimental stage slightly differs in its setup. Note that *online generalization* is mandatory. Approaches that pre-generalize the input space before learning, such as tile coding approaches (e.g. Kuvayev & Sutton, 1996), cannot solve the problem since they would learn three different models for the three stages and consequently would not be able to draw the appropriate conclusion. (It is impossible to provide an identical coding for each stage in this experiment since it is essential that no manipulanda are present during the devaluation phase.)

Without any further enhancements, ACS2 is not able to solve the task, either. To this point, the reinforcement distribution is only done during interaction with the environment. Moreover, the policy is only based on the reward prediction and the quality of the evolving environmental model. The remainder of this section shows that ACS2 can be enhanced to adapt its behavioral policy further exploiting the generalized, internal environmental model. Hereby, reinforcement is distributed internally, termed *mental acting*, or explicit anticipations influence the behavioral policy, termed *lookahead action selection*. ACS2 is able to solve the task with either anticipatory mechanism.

5.2 Mental Acting

In the *mental acting* approach, the classifier's reward prediction value r is updated internally (i.e. without environmental interaction). Anticipated events are formed in which reward predictions are evaluated and modified in the classifiers. Thus, the behavior of ACS2 is altered by executing mental actions.

Sutton (1991b) has applied a similar approach to the Dyna architecture. He showed that it is possible to adapt behavior faster in static environments, and further, achieve a faster adaptivity in dynamic environments. The environmental model was stored in a completely specialized, tabular form. The algorithm randomly updated state-action pairs by anticipating the next state and back-propagating the highest Q-value additional to the expected direct reward.

Due to the online generalized model in ACS2, the internal update process needs to be modified. First, since classifiers usually only specify parts of the perceptual attributes in their condition parts, classifiers usually predict a set of possible next states and not an exact situation-action-resulting situation triple. Second, the prediction of the next state is only valid to a degree expressed in the quality of the classifier. Finally, transitions are often represented by more than one classifier. Thus, it is necessary to assure that the relation between the classifier whose reward prediction r is updated and the classifier(s) that cause the update is reliable.

A mental action is realized by comparing effect parts of classifiers with condition parts of other classifiers. Table 1 specifies the applied *one-step mental acting* algorithm in pseudo code. The algorithm forms a link set $[L]$ that restricts the update to reliable classifier relations.

The algorithm only updates *reliable* classifiers that anticipate changes. This restricts the updates to meaningful ones and makes sure that only sufficiently stable action-outcome relations are modified. The link set $[L]$ includes all classifiers that could take place after a successful execution of classifier cl. The restriction to only those classifiers that actually explicitly specify the attributes in C that are specified in $cl.E$ is rather strong. However, this restriction proved to be necessary in the investigated tasks. Allowing more loose connections did not result in the desired learning effect. The one-step mental acting algorithm is executed after each real executed action. The number of executions is specified in the experimental runs.

In more cognitive terms, mental acting is comparable to a thought process that takes place independently of the current (outside) environment such as mental problem solving, the imagination of certain events, or even dreaming. Dreaming was recently more and more recognized as a fundamental consolidation process in learning (Stickgold, 1998) which is indeed what mental acting is doing. Mental acting causes the consolidation

of memory, that is, the consolidation of utility measures represented in reward prediction values.

Before we validate mental acting, another approach to the problem is introduced that modifies the policy determination.

5.3 Lookahead Action Selection

While *mental acting* influences action selection only indirectly, *lookahead action selection* forms explicit outcome anticipations before action execution. With respect to the theory of anticipatory behavioral control (section 2.3) this approach explicitly realizes the first point of the theory. All possible action outcome representations are formed when performing lookahead action selection. The reinforcement prediction in the outcome, then, influences action selection.

The actual algorithm is derived from the idea of a tag-mediated lookahead (Holland, 1990) and the successive implementation in CFSC2 (Riolo, 1991). While ACS2 already showed its capability of generating plans in the above section about model learning improvement, the possibility of lookahead has not been combined with the reinforcement learning procedure, yet. This is the aim of the process in this section. Instead of selecting an action according to the highest qr value in the current match set $[M]$, an action is now selected according to the currently best qr value for each possible action combined with the best qr value in the anticipated resulting state. The action selection algorithm is specified in Table 2.

First, the algorithm generates an action array of the usual values considered for action selection. Next, the result of each action is predicted, and the highest qr value in the consequent set of matching classifiers is used to update the action values in the action array. Note, as before for the best qr values, only classifiers are considered that anticipate a change. Finally, the algorithm chooses the consequent best action in the resulting action array.

In combination with the applied ϵ-greedy policy, instead of executing the best action as considered previously during exploitation, the algorithm chooses the best lookahead action for execution. For now, the algorithm is a one-step lookahead procedure. Deeper versions are possible. An animat could, for example, determine how much time it can afford to invest in a deeper action selection consideration and act accordingly.

However, the computational costs, which increase exponentially with the depth, need to be considered. In the experiments herein, we leave the question of scale-up on the side and concentrate on the general effect on behavior.

5.4 ACS2 in the Response-Effect Learning Task

To validate the two anticipatory behavior approaches, the above described environment is simulated. During the first phase, ACS2 can act upon a manipulandum and consume the possible resulting reinforcer. The consumption leads to a reinforcement of 1000, the perception of the environment without the food, and the generation of a new trial. Either lever or chain is present in each trial during this phase. In the second phase, the presence of one type of reinforcer is indicated at random. The consumption of the devalued reinforcer leads to a reinforcement of 0 while the reinforcement of the still valued reinforcer stays at 1000. After a consumption one trial ends. In the final phase, both manipulanda are present, no action leads to any effect, and the selected actions are recorded.

Environmental situations are coded by four bits. The first two bits indicate the presence of either type of reinforcer while the second two bits indicate the presence of lever or chain. The phases were executed for 204, 100, and 50 trials which approximately corresponds to the number of trials the rats experienced. Parameter settings are identical to the ones above and the curves are again averaged over 1000 runs.

Figure 8 exhibits that ACS2 is able to exploit its environmental model to simulate anticipatory controlled behavior. Regardless if mental acting, lookahead action selection, or both are applied, ACS2 consistently distinguishes the action that leads to the devalued reinforcer to the still valued one. The results show that ACS2 sufficiently generalizes the model to make the appropriate conclusions.

Additional to the confirmation of the distinction several behavioral characteristics can be observed. Similar to the rats, ACS2 decreases its distinction between the two actions during testing. In the mental acting applications with different steps, the distinction drops off faster. In the testing phase, the quality values q of the classifiers that specify the provision of one or the other reinforcer after pulling or pushing decrease under the reliability threshold since during testing no action has any effect. Thus, the mental updates do not take place anymore and the distinction between the two actions decreases faster than in the

lookahead action selection case in which anticipations are also formed with classifiers that are not reliable. In both cases, the distinction between better and worse action decreases as observed in the rats. Eventually, ACS2 does not distinguish between the two actions at all anymore since it learns that the actions do not have any effect anymore. Again, we confirmed the consistent distinction in different parameter settings for ϵ and θ_{ga} which always showed a similar distinction between the two actions. Thus, although the degree of distinction might be dependent on parameter settings, the distinction per se as well as the decrease in the distinction consistently applies throughout.

5.5 Stimulus-Response-Effect Learning Task

The stimulus-response-effect experiment was conducted with rats by Colwill and Rescorla (1990). Section 2.1 revealed the basic implications of this experiment for anticipatory controlled behavior. The experimental setup is very similar to the 1985 experiment except for the additional requirement of a stimulus distinction. Figure 2 shows the experimental setup schematically. During the first phase, an additional discriminative stimulus (noise or light) was presented that altered the response-effect pairing. During the test phase the one or the other discriminative stimulus was presented at random. Also, the first phase was altered in that at first either the one or the other manipulandum was present and later both manipulanda were present. Although with a slightly lower effect, the rats were again preferring the presumably better action during testing as shown in Figure 9.

To code the two additional discriminative stimuli, two bits are added to the previously used coding that indicate the presence of either the noise or the light stimulus. Moreover, the first phase is altered in accordance with the rat experiment executing 64 trials with either the one or the other manipulandum present and further 174 with both manipulanda present (the numbers again roughly correspond to the number of trials the rats experienced). The second phase is executed for 100 trials and the test phase for 50 trials.

The behavior of ACS2 during testing is visualized in Figure 9. Results are averaged over 1000 experiments and the parameters are set as specified above. The graphs confirm that ACS2 is able to distinguish discriminative stimuli, exploit the generalized model, and consequently adapt its behavior appropriately. In the results, the lookahead action winner method results in a much stronger effect than the mental acting

application. Due to the additional situational dependencies, mental acting is not as effective as in the first experiment since more connections can be updated.

The results confirm again the efficiency and usefulness of the evolving generalized environmental representation. Anticipatory influenced behavior is able to mimic animal behavior which would not be possible with previous mechanisms or an ALCS without processes similar to mental acting or lookahead action selection. Also the online generalization is mandatory since otherwise the knowledge transfer from the devaluation phase to the test phase would have not been possible at all. Moreover, it shows the necessary specialization of situational dependencies—the third point of the anticipatory behavioral control theory.

Both simulations show that the representation of a predictive environmental model in combination with online generalization of the model is a prerequisite for a successful simulation of rat behavior. Moreover, an additional anticipatory mechanism is necessary that influences behavior in an anticipatory fashion. In our simulations the two distinct mechanisms can cause the same behavioral effect. Whether the one, the other, both, or a different mechanism might take place in the rats is certainly not derivable from the results. However, what can be derived is that some anticipatory mechanism that influences behavior must be present.

6 Summary and Conclusions

This article provided evidence for anticipatory controlled behavior from the psychological side in exemplar animal and human experiments. Latent learning in rats suggested learning beyond the basic stimulus-response assumption in behaviorism long ago. More recently, various outcome-devaluation experiments confirmed response-outcome representations in rats. In humans, anticipations have a definite influence on response speed. Other experiments were mentioned providing evidence for anticipatory influences in reasoning, learning, attention, and preparedness.

After the provision of evidence for anticipatory influences on behavior, we suggested a basic framework of anticipatory controlled behavior. It was suggested that (1) anticipations precede any voluntary act, (2) primarily action-outcome coincidences are learned, (3) situational dependencies are learned as a secondary process, (4) needs or desires of outcomes trigger action-outcome representations, and (5) certain stimuli cause

the preparedness for action-outcome relations. The framework is partly realized in the anticipatory learning classifier system ACS2 whose performance was evaluated next. The behavioral evaluations in different rat experiments confirmed that anticipatory representations and online generalization are necessary to mimic rat behavior in various experimental setups. Not only rat behavior was mimicked but also behavior was achieved which is not possible with model-free reinforcement learning methods nor with not online generalizing model-based reinforcement learning approaches. That is, a predictive environmental model needs to be learned while interacting with the environment and the model representation needs to be generalized over the provided sensory input while interacting with the environment.

The results allow the following conclusions. (1) To enable competent adaptive behavior, explicit anticipatory influences on behavior are necessary in certain tasks. (2) To be able to realize such behavioral influences, a predictive environmental model needs to be learned online. (3) Learning of such a model should primarily form action effect relations which are conditionalized where necessary. (4) The predictive model representation needs to be generalized online over the provided perceptual input.

In the future, it is necessary to evaluate the scaling behavior of the additional anticipatory approaches pursued herein. Mental acting might be rather expensive in larger tasks and also less effective since too many relations can be updated. Prioritized updates could be helpful as for example pursued in Moore and Atkeson (1993) or Kaelbling (1993). Salient situations (such as an unexpected result) could be remembered that would further direct the internal reinforcement updates. Lookahead action selection looks only one step into the future which could be insufficient in many cases. Longer chains of lookahead, on the other hand, cause exponential computational effort. Thus, other mechanisms seem necessary to speed up the lookahead possibilities such as the formation of hierarchies in the model representation (e.g. Donnart & Meyer, 1994; Sutton, Precup, & Singh, 1999).

While ACS2 proved to be a suitable learning mechanism for the implementation of anticipatory controlled behavior, many extensions seem possible. To name a few, ACS2 should be enhanced to be able to handle stochastic environments. ACS2 should be able to ignore attributes which are not influenced by its actions as well as attributes which are irrelevant for its goals. Essentially, the current goal of ACS2,

that is, to learn a complete predictive model of the environment, should be relaxed to enable learning in more complex environments. Furthermore, more particular action and task dependent attentional processes could be included to improve and speed-up behavior. Finally, the formation of behavioral hierarchies and subprograms could allow further scalability.

With respect to adaptive behavior in anticipatory learning systems in general, it seems necessary to use anticipatory mechanisms for the realization of other cognitive processes such as attentional processes, preparedness, intentional and motivational mechanisms, as well as emotions. Anticipations should prove to be helpful for further competence in adaptive behavior as the diverse manifestations in animals and humans indicate. As a final point along this way, it still remains to be shown in which problems anticipations are actually necessary for competent adaptive behavior. This article investigated small dynamic environments in which dynamic changes demanded backward conclusions. While the demand for backward conclusions seems to be a general indicator for the utility of anticipations, future research must identify which dynamic changes demand backward conclusions, when the demand of backward conclusions actually requires anticipatory controlled behavior, and if the demand for backward conclusions is the only one in which anticipatory controlled behavior is helpful.

Acknowledgments

The authors would like to thank Wolfgang Stolzmann for his contributions to this work. Moreover, the authors would like to thank David E. Goldberg for his support as well as the whole IlliGAL lab at the University of Illinois at Urbana-Champaign including Martin Pelikan, Kumara Sastry, and others. Many thanks also to the three anonymous reviewers as well as Jason Noble for their great comments to improve this work and to make it accessible to a wider audience. Finally, the authors would like to thank their colleagues at the department of cognitive psychology at the University of Würzburg including Andrea Kiesel, Wilfried Kunde, Albrecht Sebald, Armin Stock, and Christian Stöcker. The work was supported by the German Research Foundation (DFG) under grant HO1301/4.

References

Adams, C., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, *33*(B), 109–121.

Butz, M. V. (2002). *Anticipatory learning classifier systems.* Boston, MA: Kluwer Academic Publishers.

Butz, M. V., Goldberg, D. E., & Stolzmann, W. (2000). Introducing a genetic generalization pressure to the anticipatory classifier system: Part 2 - Performance analysis. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000)*, 42–49.

Colwill, R. M., & Rescorla, R. A. (1985). Postconditioning devaluation of a reinforcer affects instrumental learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *11*(1), 120–132.

Colwill, R. M., & Rescorla, R. A. (1990). Evidence for the hierarchical structure of instrumental learning. *Animal Learning & Behavior*, *18*(1), 71–82.

Dickinson, A. (1994). Instrumental conditioning. In Mackintosh, N. (Ed.), *Animal learning and cognition* (pp. 45–79). San Diego, CA: Academic Press.

Donnart, J.-Y., & Meyer, J.-A. (1994). A hierarchical classifier system implementing a motivationally autonomous animat. *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, 144–153.

Dorigo, M., & Colombetti, M. (1997). *Robot shaping: An experiment in behavior engineering.* Cambridge, MA: MIT Press.

Drescher, G. L. (1991). *Made-up minds: A constructivist approach to artificial intelligence.* Cambridge, MA: MIT Press.

Elsner, B., & Hommel, B. (2001). Effect anticipation and action control. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 229–240.

Gérard, P., & Sigaud, O. (2001a). Adding a generalization mechanism to YACS. *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2001)*, 951–957.

Gérard, P., & Sigaud, O. (2001b). YACS: Combining dynamic programming with generalization in classifier

systems. In Lanzi, P. L., Stolzmann, W., & Wilson, S. W. (Eds.), *Advances in Learning Classifier Systems: Third International Workshop, IWLCS 2000* (pp. 52–69). Berlin Heidelberg: Springer-Verlag.

Greenwald, A. (1970). Sensory feedback mechanisms in performance control: with special reference to the ideo-motor mechanism. *Psychological Review*, *77*, 73–99.

Harle, E. (1861). Der Apparat des Willens. *Zeitschrift für Philosophie und philosophische Kritik*, *38*, 50–73.

Hoffmann, J. (1993). *Vorhersage und Erkenntnis: Die Funktion von Antizipationen in der menschlichen Verhaltenssteuerung und Wahrnehmung. [Anticipation and cognition: The function of anticipations in human behavioral control and perception.]*. Goettingen, Germany: Hogrefe.

Hoffmann, J., Sebald, A., & Stöcker, C. (2001). Irrelevant response effects improve serial learning in serial reaction time tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 470–482.

Holland, J. H. (1976). Adaptation. In Rosen, R., & Snell, F. (Eds.), *Progress in theoretical biology*, Volume 4 (pp. 263–293). New York: Academic Press.

Holland, J. H. (1990). Concerning the emergence of tag-mediated lookahead in classifier systems. In Forrest, S. (Ed.), *Emergent Computation. Proceedings of the Ninth Annual International Conference of the Center for Nonlinear Studies on Self-organizing, Collective, and Cooperative Phenomena in Natural and Artificial Computing Networks. A special issue of Physica D.*, Volume 42 (pp. 188–201). Elsevier Science Publishers.

Hommel, B. (1996). The cognitive representation of action: Automatic integration of perceived action effects. *Psychological Research*, *59*, 176–186.

Hommel, B. (1998). Perceiving ones own action - and what it leads to. In Jordan, J. S. (Ed.), *Systems theory and apriori aspects of perception* (pp. 143–179). Amsterdam: North Holland.

James, W. (1981 (orig.1890)). *The principles of psychology*, Volume 2. Cambridge, MA: Harvard University Press.

Kaelbling, L. P. (1993). *Learning in embedded systems*. Cambridge, MA: MIT Press.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–258.

Kunde, W. (2001). Response-effect compatibility in manual choice reaction tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 387–394.

Kuvayev, L., & Sutton, R. S. (1996). Model-based reinforcement learning with an approximate, learned model. *Proceedings of the Ninth Yale Workshop on Adaptive and Learning Systems*, 101–105.

Lanzi, P. L., Stolzmann, W., & Wilson, S. W. (Eds.) (2000). *Learning classifier systems: From foundations to applications*. Berlin Heidelberg: Springer-Verlag.

Lotze, H. (1852). *Medizinische psychologie oder physiologie der seele*. Leipzig: Weidmann'sche Buchhandlung.

Moore, A. W., & Atkeson, C. (1993). Prioritizes sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, *13*, 103–130.

Münsterberg, H. (1889). *Beiträge zur Experimentalpsychologie. Heft 1*. Greiburg i.B.: J.C.B. Mohr.

Pashler, H., Johnston, J. C., & Ruthruff, E. (2001). Attention and performance. *Annual Review of Psychology*, *52*, 629–651.

Pearce, J. M. (1997). *Animal learning and cognition (2nd edition)*. Hove: Psychology Press.

Prinz, W. (1990). A common coding approach to perception and action. In Neumann, O., & Prinz, W. (Eds.), *Relationships between perception and action* (pp. 167–201). Berlin Heidelberg: Springer-Verlag.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology*, *9*, 129–154.

Rescorla, R. A. (1990). Evidence for an association between the discriminative stimulus and the response-outcome association in instrumental learning. *Journal of Experimental Psychology: Adaptive Behavior Processes*, *16*(4), 326–334.

Rescorla, R. A. (1991). Associative relations in instrumental learning: The eighteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology*, *43*(B), 1–23.

Rescorla, R. A. (1995). Full preservation of a response-outcome association through training with a second outcome. *Quarterly Journal of Experimental Psychology*, *48*(B), 252–261.

Riolo, R. L. (1991). Lookahead planning and latent learning in a classifier system. *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, 316–326.

Roitblat, H. L. (1994). Mechanism and process in animal behavior: Models of animals, animals as models. *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, 12–21.

Rosen, R. (1985). *Anticipatory systems.* Oxford, UK: Pergamon Press.

Rosen, R. (1991). *Life itself.* New York: Columbia University Press.

Schubotz, R. I., & von Cramon, D. Y. (2001). Functional organization of the lateral premotor cortex. fMRI reveals different regions activated by anticipation of object properties, location and speed. *Cognitive Brain Research*, *11*, 97–112.

Stickgold, R. (1998). Sleep: Off-line memory reprocessing. *Trends in Cognitive Sciences*, *2*(12), 484–492.

Stock, A., & Hoffmann, J. (2002). Intentional fixation of behavioral learning or how R-E learning blocks S-R learning. *European Journal of Cognitive Psychology*, *14*(1), 127–153.

Stolzmann, W. (1997). *Antizipative Classifier Systems [Anticipatory classifier systems].* Aachen, Germany: Shaker Verlag.

Stolzmann, W. (2000). An introduction to anticipatory classifier systems. In Lanzi, P. L., Stolzmann, W., & Wilson, S. W. (Eds.), *Learning Classifier Systems: From Foundations to Applications* (pp. 175–194). Berlin Heidelberg: Springer-Verlag.

Stolzmann, W., Butz, M. V., Hoffmann, J., & Goldberg, D. E. (2000). First cognitive capabilities in the anticipatory classifier system. *From Animals to Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*, 287–296.

Sutton, R., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, *112*, 181–211.

Sutton, R. S. (1991a). Dyna, an integrated architecture for learning, planning, and reacting. *Working Notes of the 1991 AAAI Spring Symposium on Integrated Intelligent Architectures*, 151–155.

Sutton, R. S. (1991b). Reinforcement learning architectures for animats. *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, 288–296.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

Tani, J. (1996). Model-based learning for mobile robot navigation from the dynamical systems perspective. *IEEE Transactions. System, Man and Cybernetics (Part B), Special Issue on Learning Autonomous Systems*, *26*(3), 421–436.

Thislethwaite, D. (1951). A critical review of latent learning and related experiments. *Psychological Bulletin*, *48*(2), 97–129.

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies.* New York: Macmillan.

Tolman, E. C. (1932). *Purposive behavior in animals and men.* New York: Appleton.

Tolman, E. C. (1949). There is more than one kind of learning. *Psychological Review*, *5b*, 144–155.

Tolman, E. C., & Honzik, C. (1930). Introduction and removal of reward, and maze performance in rats. *University of California, Publications in Psychology*, *4*, 257–275.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards.* Doctoral dissertation, King's College, Cambridge, UK.

Whitehead, S. D., & Ballard, D. H. (1991). Learning to perceive and act. *Machine Learning*, *7*(1), 45–83.

Wilson, S. W. (1991). The animat path to AI. *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, 15–21.

Wilson, S. W. (1995). Classifier fitness based on accuracy. *Evolutionary Computation*, *3*(2), 149–175.

Wilson, S. W. (1998). Generalization in the XCS classifier system. *Genetic Programming 1998: Proceedings of the Third Annual Conference*, 665–674.

Witkowski, C. M. (1997). *Schemes for learning and behaviour: A new expectancy model.* Doctoral dissertation, Department of Computer Science, Queen Mary Westfield College, University of London.

Witkowski, C. M. (2000). The role of behavioral extinction in animat action selection. *From Animals to*

Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior, 177–186.

Ziessler, M. (1998). Response-effect learning as a major component of implicit serial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 962–978.

Ziessler, M., & Nattkemper, D. (2001). Learning of event sequences is based on response-effect learning: Further evidence from serial reaction task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 595–613.

Table 1: The *One-Step Mental Acting* algorithm in pseudo code

One-Step Mental Acting:

1 Choose at random a reliable classifier cl from the populaton $[P]$ that predicts a perceptual change.
2 Form a link set $[L]$ of classifiers whose conditions specify all attributes in $cl.E$ and don't mismatch the explicitly non-changing attributes in cl.
3 Back-propagate the best qr value of a classifier in $[L]$ updating the reward prediction $cl.r$, i.e. $cl.r = cl.r + \beta(cl.ir + \gamma \max_{c \in [L]}(c.q \cdot c.r) - cl.r)$

Table 2: Algorithmic description of the *Choose Best Lookahead Action* algorithm

Choose Best Lookahead Action:

1 Generate action array AA that specifies the best qr value for each possible action.
2 for each possible action a
3 choose classifier cl_a with the highest quality q among all classifiers that specify a perceptual change and action a
4 if there is a classifier cl_a
5 modify $AA[a]$ with the highest qr classifier c that matches the prediction of cl_a and specifies a change, i.e. $AA[a] \leftarrow (AA[a] + cl_a.q \cdot \gamma \cdot c.q \cdot c.r)/(1 + cl_a.q \cdot \gamma)$
6 Choose best action according to $[AA]$

Figure 1: In Colwill and Rescorla (1985), rats are able to either press a lever or pull a chain (R_1, R_2) that leads to either food pellets or sucrose. After the devaluation of one outcome (O_1), the action is preferred, that previously led to the other outcome (O_2). The result is not explainable with a stimulus-response approach.

Figure 2: Colwill and Rescorla (1990) show that some form of situational dependent $R - O$ relations are learned by rats. After teaching the rats different $S - R - O$ relations (light or noise in combination with pressing a lever or pulling a chain leads to food pellet or sucrose) and the devaluation of one outcome, the action is preferred that dependent on the situation (light or noise) previously led to the other outcome.

Figure 3: The theory of anticipatory behavioral control emphasizes the initial bias towards learning action-effect relations. The consideration of situational dependencies is regarded as a secondary process.

Figure 4: During one agent/environment interaction, ACS2 forms a match set representing the predictive knowledge with respect to the current perceptions. Next, it generates an action set representing the knowledge about the consequences of the chosen action in the given situation. Classifier parameters are updated by RL and ALP. Moreover, new classifiers might be added and old classifiers might be deleted by genetic generalization and ALP.

Figure 5: In the experiment by Rescorla (1990), rats learn $R - O$ relations (pressing a lever or pulling a chain leads to food pellets or sucrose) that depend on discriminative stimuli (noise, light, or tone). In phase two, the associations $R_1 - O_1$ and $R_2 - O_2$ are extinct by providing no more food in the light condition. Consequently, dependent on the presented stimulus, the rats prefer to execute that action in the test phase that led to the $R-O$ relation in phase one that was not extinct in phase two. Thus, hierarchical $S-(R-O)$ relations influence behavior.

Figure 6: ACS2 exhibits behavior similar to the behavior observed in rats in the simulation of the Rescorla (1990) experiment. Depicted are the number of actions executed in the beginning, after 60 steps, and after 140 steps in the test phase by ACS2 and the mean action execution in the beginning and in the second half of the experiment by the rats. "Different" refers to the action that previously led to the not extinct $R - O$ relation, while "same" refers to the extinct $R - O$ relation. "Other" refers to the execution of an "eating" or "do nothing" action in ACS2 during testing. "ITI" refers to the mean response of pulling and pushing during inter trial intervals. Three similarities between the rat behavior and the behavior of ACS2 can be observed: (1) The "different" action is preferred. (2) The preference decreases over the test phase (difference between "different" and "same"). (3) The frequency of acting upon the manipulanda decreases (in ACS2: since the "other" actions are increasingly executed; in the rats: decrease in mean response per minute).

Figure 7: In all three different settings, rats preferred the action that previously led to the still valued reinforcer to the one that led to the now less valued one in the Colwill and Rescorla (1985) experiments. In the first and second setting, one reinforcer was devalued by pairing its consumption with $LiCl$ while in the last experiment one reinforcer was sated.

Figure 8: In the simulation of the Colwill and Rescorla (1985) experiment, ACS2 is able to exploit its online generalized environmental model for an adaptive behavior beyond model-free RL and not online generalizing model-based RL architectures. Regardless if lookahead action selection or mental acting is applied, ACS2 prefers that action that previously led to the not devalued outcome.

Figure 9: The results in the simulation of the stimulus-response-effect experiment show that ACS2 is able to further adapt its behavior differentiating between different stimuli similar to the differentiation observed in rats. Again, adaptive behavior beyond model-free RL approaches or not online generalizing model-based RL architectures is achieved.

Training	Devaluation	Test
$R_1 \rightarrow O_1$		
	$O_1 \rightarrow LiCl$	$R_1 < R_2$
$R_2 \rightarrow O_2$		

Figure 1:

Training	Devaluation	Test
$S_1 : R_1 \rightarrow O_1, R_2 \rightarrow O_2$		$S_1 : R_1 < R_2$
	$O_1 \rightarrow LiCl$	
$S_2 : R_1 \rightarrow O_2, R_2 \rightarrow O_1$		$S_2 : R_1 > R_2$

Figure 2:

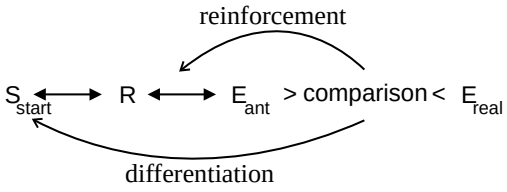


Figure 3:

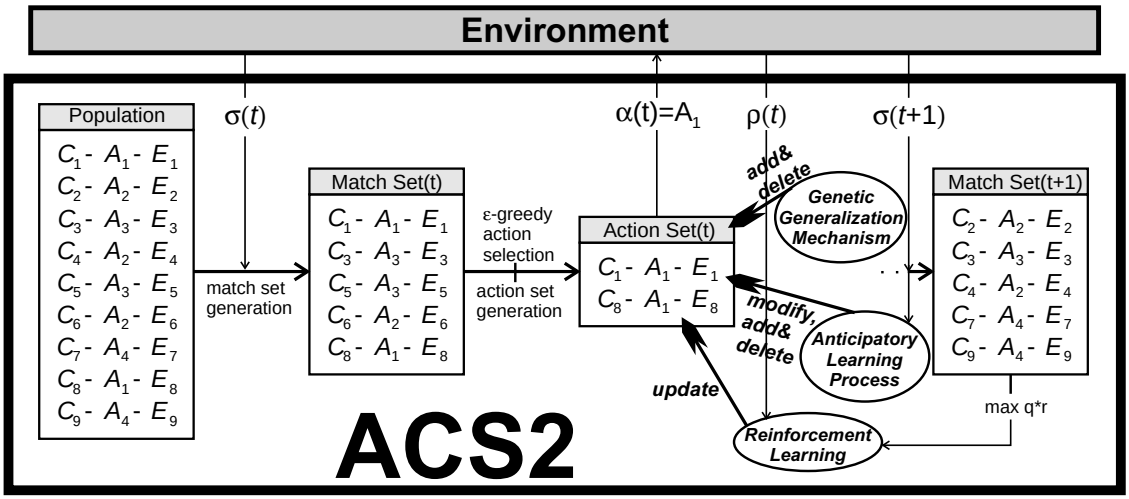
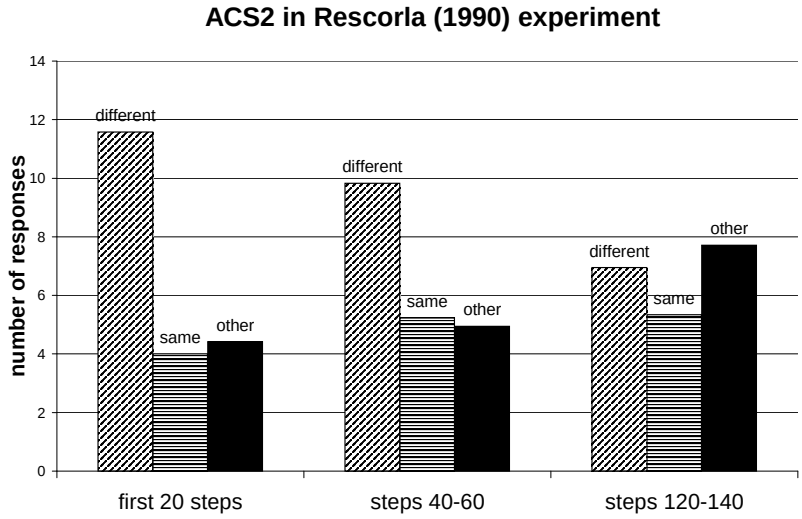


Figure 4:

Training	Extinction	Test
$N : R_1 \rightarrow O_1, R_2 \rightarrow O_1$		$N : R_1 < R_2$
$L : R_1 \rightarrow O_1, R_2 \rightarrow O_2$	$L : R_1-,\ R_2-$	
$T : R_1 \rightarrow O_2, R_2 \rightarrow O_2$		$T : R_1 > R_2$

Figure 5:


ACS2 in Rescorla (1990) experiment
number of responses
0
2
4
6
8
10
12
14
different
same
other
different
same
other
different
same
other
first 20 steps
steps 40-60
steps 120-140

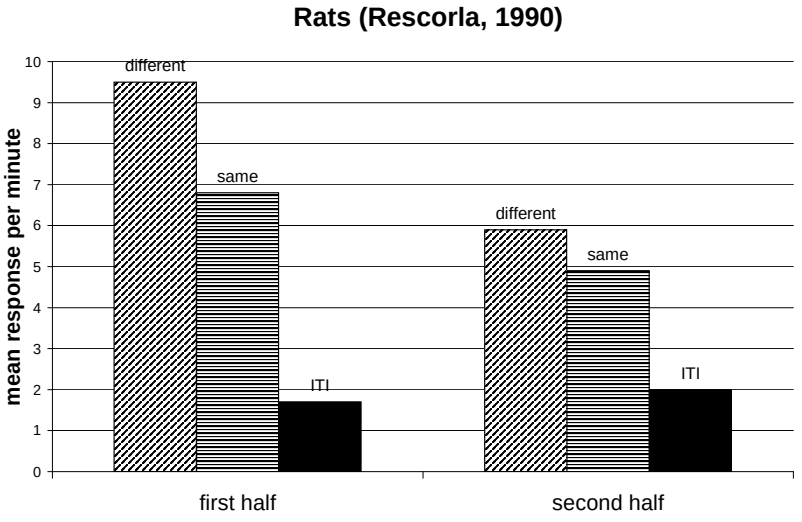

Rats (Rescorla, 1990)
mean response per minute
0
1
2
3
4
5
6
7
8
9
10
different
same
ITI
different
same
ITI
first half
second half

Figure 6:

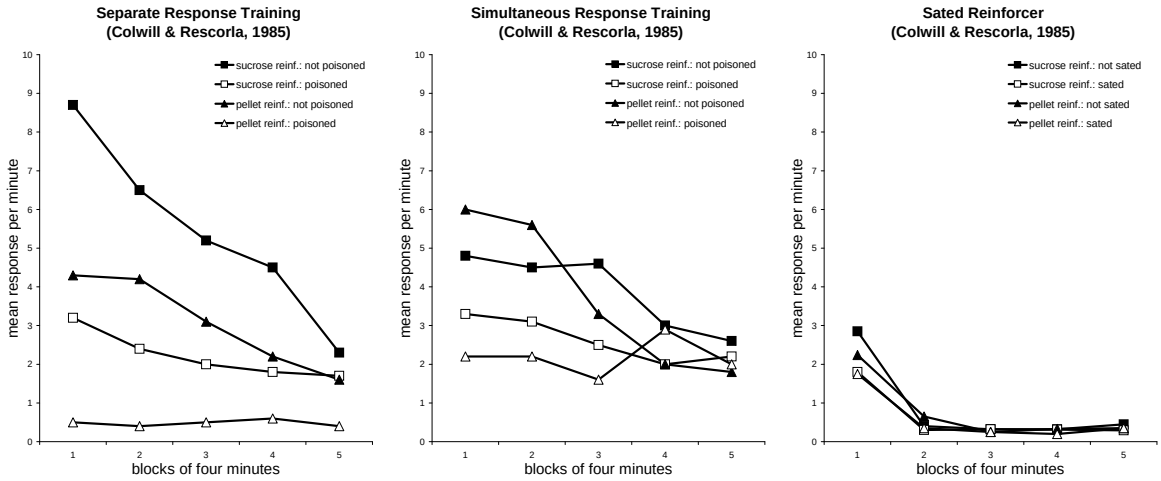


Figure 7:

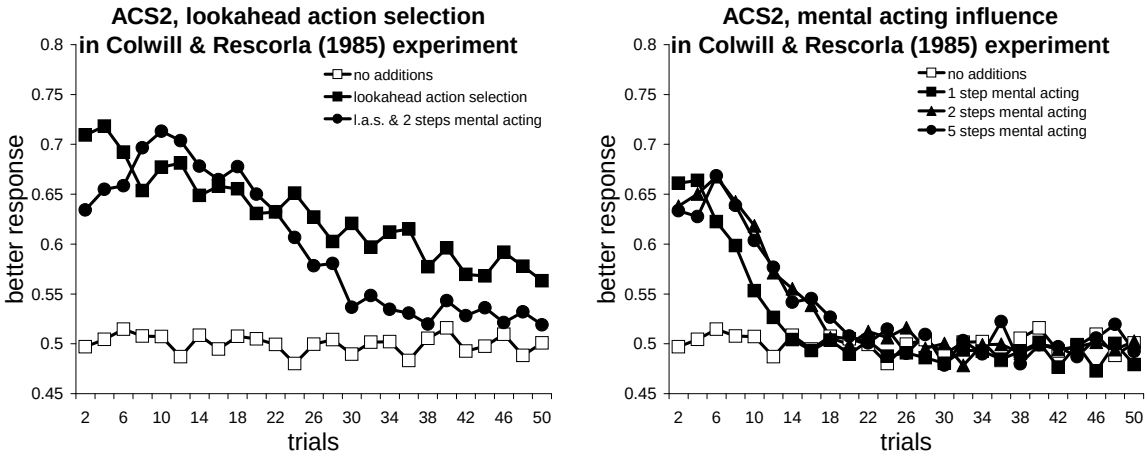


Figure 8:

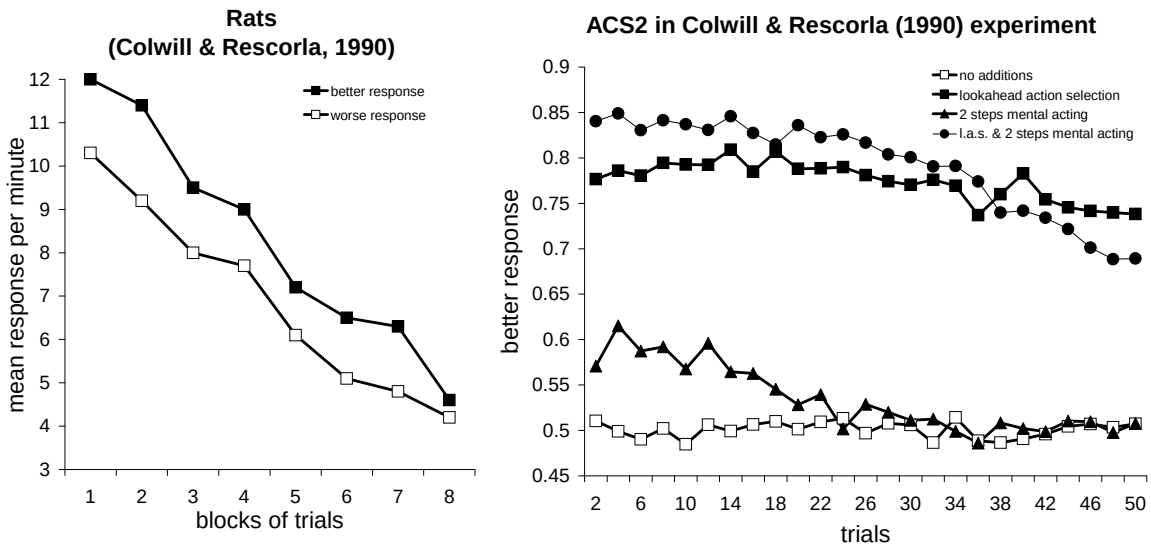

Rats
(Colwill & Rescorla, 1990)
mean response per minute
blocks of trials
better response
worse response
ACS2 in Colwill & Rescorla (1990) experiment
better response
trials
no additions
lookahead action selection
2 steps mental acting
l.a.s. & 2 steps mental acting

Figure 9:

Figure 10: Martin V. Butz

Martin V. Butz

Marin Butz is a PhD student in computer science at the University of Illinois at Urbana-Champaign. He received his Diploma in computer science from the University of Würzburg in 2001. Martin Butz is working at the Illinois Genetic Algorithms Laboratory (IlliGAL) as well as at the department of cognitive psychology at the University of Würzburg.

Martin Butz's major research interest lies in the study of anticipatory learning and anticipatory behavior. Moreover, he is working on the relation of these mechanisms to general learning theories in machine learning as well as to cognitive mechanisms yielding competent adaptive behavior.

Figure 11: Joachim Hoffmann

Joachim Hoffmann

Joachim Hoffmann is professor of cognitive psychology at the University of Würzburg. He received his Ph.D in 1978 from Humboldt University in Psychology.

Dr. Joachim Hoffmann's actual research interests center on the interplay between cognition and behavioral control. He is especially interested in examining of how cognitive processes like perception, attention, memory, and learning are shaped by the function they serve in the selection, initiation, and execution of voluntary behavior.