Affect and Learning: a computational analysis

Broekens, D.J.

3

Affect and Exploration

Affect-Controlled Exploration is Beneficial to Learning

Recent studies show that affect influences and regulates learning. We report on a computational study investigating this. We simulate affect in a probabilistic learning agent and dynamically couple affect to its action-selection mechanism, effectively controlling exploration versus exploitation behavior. The agent's performance on two types of learning problems is measured. The first consists of learning to cope with two alternating goals. The second consists of learning to prefer a later larger reward (global optimum) to an earlier smaller one (local optimum). Results show that, compared to the non-affective control condition, coupling positive affect to exploitation and negative affect to exploration has several important benefits. In the Alternating-Goal task, it significantly reduces the agent's "goal-switch search peak". The agent finds its new goal faster. In the second task, artificial affect facilitates convergence to a global instead of a local optimum, while permitting to exploit that local optimum. Our results illuminate the process of affective influence on learning, and furthermore show that both negative affect and positive affect can be beneficial to learning. Further, our results provide evidence for the idea that negative affect is related to less selective decisions while positive affect is related to more selective decisions.

3.1 Introduction

As we have seen in Chapter 1 and 2, emotions influence thought and behavior in many ways. In this chapter we focus on the influence of affect on learning and adaptation. The main question we address here is: how is an agent's learning performance influenced if artificial affect is used to control exploration versus exploitation. Based on findings from the affect-cognition literature (Craig, Graesser, Sullins & Gholson, 2004; Dreisbach & Goschke, 2004; Rose, Futterweit & Jankowski, 1999) as discussed in Chapter 2, we hypothesize two types of relations between affect and exploration. The first type relates positive affect to exploitation, and negative affect to exploration. The second type uses the inverse relation of the first type, i.e., positive affect relates to exploration while negative affect relates to exploitation. We contrast these two dynamic settings to a non-affective control group of agents that use a static amount of exploration.

We investigate the relation between affect and learning with a self-adaptive agent in a simulated grid world. The agent acts in the grid world—in our case a simulated maze that represents a psychological task—and builds a model of that world based on perception of its surroundings and received rewards. Our agent

autonomously influences its action-selection mechanism—the agent's mechanism that proposes next actions based on the learned model. The agent uses artificial affect, as defined in Chapter 2, to control the randomness of action selection. This enables the agent to autonomously vary between exploration and exploitation.

Our agent learns (adapts) using a simple form of Reinforcement Learning. The agent learns by constructing a Markov Decision Process (MDP), of which the state-value pairs are learned using a mechanism based on model-based Reinforcement Learning (Kaelbling, Littman & Moore, 1996). We investigate the hypothesized relations between affect and exploration using two different learning tasks (modeled as discrete grid worlds). In the first task the agent has to cope with a sudden switch from an old goal in one arm of a two-armed maze to a new goal in the other arm. We call this task the Alternating-Goal task. The second task consists of learning to prefer a later larger reward (global optimum) to an earlier smaller one (local optimum). We call the second task the "Candy task"; candy represents the local optimum being closest to the agent's starting position, while food represents the global optimum being farther away from its starting position.

From a learning and adaptation point of view, these tasks represent two significant problems for an agent. The Alternating-Goal task exposes an agent to a changing set of goals. The agent has to modify its behavior in order to reflect a change in this set of goals. It has to be flexible enough to give up on an old goal and learn a new one, while at the same time it has to be persistent enough to continue trying an active goal in order to actually learn the path to the goal (Dreisbach & Goschke, 2004). In other words, to cope with alternating goals, the agent has to decide when to explore its environment and when to exploit its knowledge; a.k.a. the exploration-exploitation problem or tradeoff (Kaelbling, Littman, & Moore, 1996). In our task, failure to solve this problem results in huge goal-switch cost (if the agent does not explore the environment after the goal-switch has taken place) and/or slow/unstable convergence (if, after exploration, the agent does not exploit its learned new model of the environment).

The Candy task represents searching for a global optimum, while exploiting a newly found local optimum. This ability is important for adaptive agents as it enables them to survive with the knowledge they have, while trying to find better alternatives. Failure to do so results in getting stuck in local optima or slow convergence. This again represents a tradeoff between persistence and flexibility, but different from the tradeoff in the first task. Now, the agent has to autonomously decide that the current goal *might* not be good enough and search for a better goal. In contrast, in the previous task the old goal attractor (high reward) is removed and the agent should react to this by searching for a new goal.

In this study we use artificial affect as defined in Chapter 2, that is, artificial affect is a measure for how well the agent is doing compared to what it is used to, based on an analysis of the difference between a long-term and a short-term reinforcement signal average. In the next section we explain our experimental method, i.e., how we implemented the two different relations between affect and action selection mentioned earlier, the grid-world setup, the tasks, the agent's learning mechanism and our experimental setup. In Section 3.3 we present experimental results. Section 3.4 discusses these results in a broader context.

3.2 Method

To investigate the influence of affect-controlled exploration, we did experiments in two different simulated mazes. Each maze represents a task, and we compared affect-controlled dynamic exploration to several control conditions with static amounts of exploration.

3.2.1 Learning Environment.

The first task is a two-armed maze with a potential goal at the end of each arm (Figure 3.1a). This maze is used for the Alternating-Goal task, i.e., coping with two alternating goals: find food or find water (only one goal is active during an individual trial, goal reward r = +2.0). The second maze has *two active* goal locations (Figure 3.1b). The nearest goal location is the location of the candy (i.e., a location with a reward r = +0.25), while the farthest goal location is the food location (r = +1.0). This maze is used for the Candy task. The walls in the mazes are "lava" patches, on which the agent can walk, but is discouraged to do so by a negative reinforcement (r = –1.0).

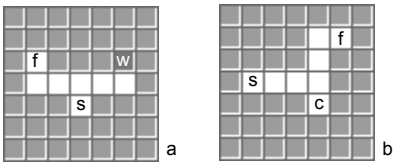

Figure 3.1. Mazes used in the experiments; (a) the Alternating-Goal task, (b) the Candy task; the 's' denotes the agent's starting position, 'f' is food, 'c' is candy and 'w' is water.

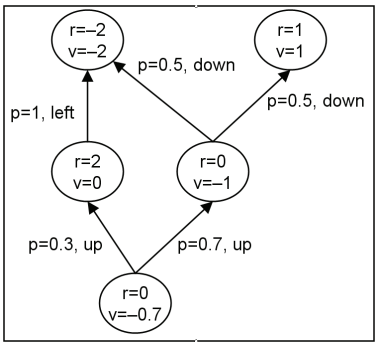

Figure 3.2. Example of a Markov Decision Process. Nodes are agent-environment states. Edges are actions with the probability *p* that executing the action results in the state to which the edge points. Nodes contain rewards (*r*; local reinforcement) and values (*v*; future reinforcement). In this example γ=1 (see text).

The agent learns by acting in the maze and by perceiving its direct environment using an 8-neighbor and center metric (i.e., it senses its eight neighboring locations and the location it is at). An agent that arrives at a goal location is again placed at its starting location. Agents learn a probabilistic model of the actions and their values possible in the world. Mathematical details of this process follow; however, the most important part of our method is explained in Section 3.2.2. Agents start with an empty model of the world and construct a Markov Decision Process (MDP) as usual (i.e., a perceived stimulus is a state s in the MDP, and an action a leading from state s_1 to s_2 is an edge in the MDP; see Figure 3.2; for details see Sutton & Barto, 1998). The agent counts how often it has seen a certain state s, $N(s)$. It uses this statistic to learn the value function, $V(s)$ (comparable to model-based Reinforcement Learning, see, e.g., Kaelbling et al., 1996). This function learns to predict a cumulative future reward for every observed state. This $V(s)$ is learned in the following way. A reward function, $R(s)$, learns to predict the local reward of a state:

$$R(s) \leftarrow R(s) + \alpha \cdot (r - R(s)) \qquad (3.1)$$

This learned reward is used in the value function $V(s)$:

$$V(s) \leftarrow \gamma \sum_i \left(\frac{N(s_{a_i})}{\sum_j N(s_{a_j})} V(s_{a_i}) \right) + R(s) \qquad (3.2)$$

So, a state s has two reinforcement-related properties: a learned reward value $R(s)$ and a value $V(s)$ that incorporates predicted future reward. The $R(s)$ value converges to the local reward for state s with a speed proportional to the learning rate α. The final value of s, $V(s)$, is updated based on $R(s)$ and the weighted predicted rewards of the next states reachable by actions a_i. In Reinforcement

Learning, the discount factor γ defines how important future versus current reward is in the construction of the value function, $V(s)$. If the discount factor, γ, is equal to 1, future reward is important (no discount), while $\gamma = 0$ means that only local reward is important for the construction of the value of a state as expressed by $V(s)$. In the Alternating-Goal task the learning rate α and discount factor γ are respectively 1.0 and 0.7, and in the Candy task respectively 1.0 and 0.8.

3.2.2 Modeling Action Selection.

Most relevant to the current study is that our agent uses the Boltzmann distribution to select actions based on learned values of predicted next states. This function is often used in Reinforcement Learning and is particularly useful as it enables both exploration and exploitation:

$$p(a) = \frac{\exp[\beta \times V(s_a)]}{\sum_{i=1}^{|A|} \exp[\beta \times V(s_{a_i})]} \qquad (3.3)$$

Here, $p(a)$ is the probability that the agent chooses action a, and $V(s_a)$ is the value of a next state predicted by action a. $|A|$ is the size of the set A containing the agent's potential actions[1]. Importantly, the *inverse* temperature parameter β determines the randomness of the distribution. The larger the β the more this distribution adopts a greedy selection strategy (thus little variation in deciding what action to perform in a certain state). If β is zero the distribution function adopts a uniform random selection strategy, regardless of the predicted reward values (thus high variation in deciding what next action to perform in a certain state).

De facto, the β parameter can be used to vary the adaptive agent's processing strategy between exploration and exploitation. Note that we define exploration as generating new learning experiences by selecting actions that are non-optimal according to the current model the agent has learned, while exploitation is defined as selecting optimal actions according to the currently learned model. Therefore, if we assume, for simplicity, that the model is a tree with the agent's starting state as root and edges as different actions to different next states, exploration generates different paths through the tree at different runs, while exploitation

[1] Note that for notational simplicity we assume that an action in one state leads to a determined next state, i.e., the world is deterministic and completely observable. However, our first world is not deterministic as we introduce a for the agent non-predictable goal-switch.

retries the same paths at different runs. In a lazy value propagation mechanism as ours, exploration is needed to find solutions, while exploitation is needed to internalize solutions. Exploitation thus models animal learning by repetition, while exploration models animal search.

Key in our study is that our agent uses its artificial affect e_p to control its β parameter. Affect directly and dynamically controls exploration versus exploitation. This approach is compatible with viewing emotion as a mechanism for meta-learning (Doya, 2000; Doya, 2002; Schweighofer & Doya, 2003).

3.2.3 Type-A: Positive Affect Relates to Exploitation

To investigate how affect can influence exploration versus exploitation, we hypothesize the following two relations. First, type-A agents model positive affect related to increased exploitation:

$$\beta = e_p \times (\beta_{\max} - \beta_{\min}) + \beta_{\min} \qquad (3.4)$$

If affect e_p increases to 1, β increases towards β_{max} and as e_p decreases to 0, β consequently decreases towards β_{min}. So positive affect results in more exploitation, while neutral and negative affect results in more exploration, as suggested by the study by Rose et al. (1999), detailed in Chapter 2. This is also compatible with the idea that positive mood relates to top-down processing (Gasper & Clore, 2002), i.e., in our case to the agent using its learned model to control its behavior. A selective mode of action selection uses this model to drive behavior, while a less selective mode could be said to use more diverse behaviors (whether or not this also models bottom-up processing is unclear).

3.2.4 Type-B: Negative Affect Relates to Exploitation

The second relation is the inverse of the first one. Type-B agents thus model positive affect related to increased exploration. Positive affect favors detaching actual behavior from existing goals (as suggested by the results of the study by Dreisbach and Goschke (2004):

$$\beta = (1 - e_p) \times (\beta_{\max} - \beta_{\min}) + \beta_{\min} \qquad (3.5)$$

As affect e_p increases to 1, β decreases towards β_{min} and as e_p decreases to 0, β consequently increases towards β_{max}. So, positive affect results in more exploration, while negative affect results in more exploitation.

Of course, cognitive set-switching and attention are not equivalent to learning. Both are a precursor to learning, specifically explorative learning. Divided attention and flexible set-switching enable an individual to faster react to novel situations by favoring processing of many external stimuli. So, in the study by Dreisbach and Goschke (2004) *positive* affect facilitated exploration, as it helped to remove bias towards solving the old task thereby enabling the subject to faster adapt to the new task. However, in the study by Rose, Futterweit and Jankowski (1999) neutral affect facilitated exploration as it related to defocused attention.

3.2.5 Experimental Procedure

To investigate the influence of affect-controlled exploration, our experiments are repeated with agents of type-A and type-B as well as a control condition of agents that use static levels of exploration versus exploitation (fixed β). In the Alternating-Goal task agents first have to learn goal one (food). After 200 trials the reinforcement for food is set at $r = 0.0$, while the reinforcement for water is set at $r = +2.0$. The water is now the active goal location (so an agent is only reset at its starting location if it reaches the water). This reflects a task-switch, of which the agent is unaware. It has to search for the new goal location. After 200 trials, the situation is set back; i.e., food becomes the active goal. This is repeated 2 times resulting in 5 phases, i.e., initial learning of food goal (phase 0), then water (phase 1), food (2), water (3), and finally food (4). This (5 phases, a total of 1000 trials) represents 1 run. We repeated runs to reach sufficient statistical power. All Alternating-Goal task results are based on 800 runs, while Candy task results are based on 400 runs. During a run, we measured the number of steps needed to get to the goal (steps needed to end one trial), resulting in a learning curve when averaged over the number of runs. We also measured the average β (resulting in an "exploration-exploitation" curve), and we measured the quality of life (QOL) of the agent (measured as the sum of the rewards received during one trial). The problem for the agent is to exploit the goal but at the same time "survive" a goal switch, i.e., keep the switch-cost as low as possible. So, the learning curve of the trials just after the task-switch indicate how flexible the agent is.

The setup of the Candy task experiment is simpler, and we measured the same (steps, β and QOL). The agent has to learn to optimize reward in the Candy maze. The problem for the agent is to (1) exploit the local reward (candy), but at the same time (2) explore and then exploit the global reward (food). This relates to opportunism, an important ability that should be provided by an action-selection mechanism (Tyrell, 1993). Average QOL curves will thus show to what extent an agent has learned to exploit the global reward.

Our independent variable is the type of exploration-exploitation control. We have several different settings of type-A ("*dyn*" in Figure 3.3-3.11) and type-B ("*dyn inv*" in Figure 3.3-3.11) affect-controlled exploration. For example, "AG dyn 3-6" means that the agent was tested in the Alternating-Goal task using affect controlled exploration of type-A (positive affect relates to exploitation) with exploration-exploitation varying respectively between β_{min}=3 and β_{max}=6 (see also Figure 3.3). The artificial affect parameters *star* and *ltar* defining the short-term period and the long-term period over which artificial affect is measured were set at 50 and 375 steps respectively. As a control condition we used agents with different static amounts of exploration ("*static*" in Figure 3.3-3.11). High static β values model low exploration and high exploitation while low values denote high exploration and low exploitation. The legend of Figure 3.7 shows all different agents used in the Alternating-Goal task. Figures 3.3-3.6 show relevant subsets of these agents. The legends of Figures 3.8-3.11 show all agents used in the Candy task, excluding static agents with β=5 and β=7. The results from these two agents did not add anything to the analysis and are therefore omitted.

3.3 Results

We now discuss the results of the experiments. A discussion in a broader context is presented in Section 3.4.

3.3.1 Experiment 1: Alternating-Goal Task

Our main finding is that type-A (positive affect relates to exploitation, negative to exploration) results in the lowest switch cost between different goals, as measured by the number of steps taken *at the trial in which the goal switch is made* (Figure 3.7). This is an important adaptation benefit. As shown, all goal-switch peaks (phases 1-4) of the 4 variations of type-A (i.e., dotted lines labeled AG dyn 3-6, 3-7, 3-9 and 2-8) are smaller than the peaks of the control (straight lines labeled AG static 3, 4, 5, 6 and 7) and type-B (i.e., striped lines labeled AG dyn inv 3-6, 3-7, 3-9 and 4-9). Initial learning (phase 0) is marginally influenced by affective feedback and by static β settings (Figure not shown). Closer investigation of the first goal switch (trial 200; phase 1; Figure 3.4) shows that the trials just after the goal-switch also benefit considerably from type-A. When we computed for all settings an average peak for trail 200, 201 and 202 together, and compared these averages statistically, we found that type-A performs significantly better (p<0.001 for all comparisons, Mann-Whitney, n=800). Closer investigation of the fourth goal-switch (trial 800, phase 4; Figure 3.5), reveals a different picture. Only the trial in which the goal is switched benefits significantly from type-A (p<0.001 for all comparisons except those mentioned shortly, Mann-Whitney, n=800).

Comparison between type-A (AG dyn 3-6, 3-7 and 3-9) and AG static β=6 showed significant smaller peaks for type-A with $p<0.01$, $p<0.05$, and $p<0.01$ respectively. So it seems that a high static amount of exploration performs slightly better at later goal switches but worse at earlier goal-switches as compared to affective control over exploration. One reason for this is that the agent has built up a very good model of both arms of the maze in these later phases. This means that in later phases, less exploration is needed anyway, because the agent only needs to relearn to take the right choice at the T-junction, but not learn the new arm in the maze. This limits the potential gain of affective control. This explanation is supported by the peak curves in Figure 3.7. Here, higher β values perform worse than lower at the peaks of earlier phases but better at the peaks of later phases. Note that for the first phase, this is also true, but as we plot only the first trial after the goals-switch in Figure 3.7 this is not shown (it *is* shown in Figure 3.4, where we detail the peak of the first phase, high β values show higher peaks than do low β values).

All other comparisons between peaks revealed significantly ($p<0.001$) smaller peaks for type-A. This effect is most clearly shown for the peaks of phase 3 and 4, where the peak-height difference between type-A peaks and static peaks is a factor 1.25 to 2. This means that the type-A model of affective control of action selection can result in up to a 2-fold decrease of search investment needed to find a new goal. As expected, the smallest difference between control and type-A is when β is small (3 or 4) in the control condition (small β = much exploration = less tied to old goal). However, small β's have a classical downside: less convergence (Figure 3.6). The agent is less able to exploit its model of the world and thus does not learn the solution well, while type-A curves in Figure 3.6 show that the agent does converge to the minimum number of steps needed to get to the goal (i.e., 4 steps).

For completeness we show the β curves for the complete phase 1 of the control group agents, one type-A agent and one type-B agent (Figure 3.3). These curves confirm the expected β dynamics. For type-A, the goal switch induces high exploration (β near β_{min}) due to the lack of reinforcement ("it is going worse than expected"), after which β quickly moves up to β_{max}, and then decays to average. For type-B this behavior is exactly the opposite.

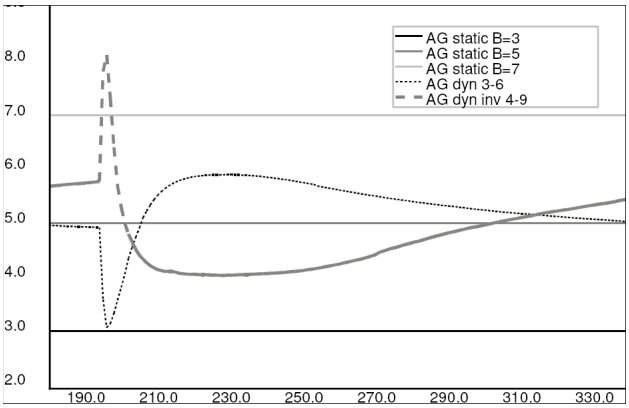

Figure 3.3. Alternating-Goal task; plot of the mean Boltzmann β for phase 1 (n=800). High β represents exploitation, low β represents exploration. The values of β for three static and two dynamic agents are shown. In all graphs, the trials are on x-axis, and means are based on the 5-95% percentile. Here, mean β is on the y-axis.

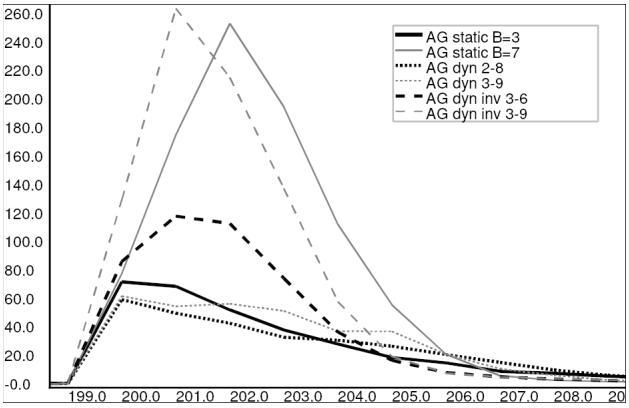

Figure 3.4. Alternating-Goal task; mean learning curves for phase 1 peak (n=800). The mean number of steps (y-axis) needed to find the goal is plotted per trial for two static, two dynamic and two inverse-dynamic agents (see text for explanation).

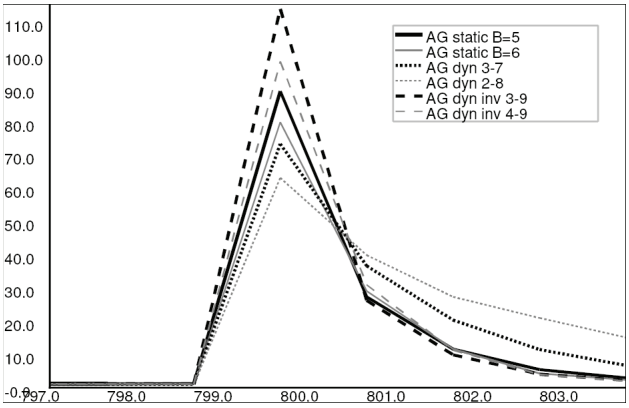

Figure 3.5. Alternating-Goal task; mean learning curves for phase 4 peak (n=800). The mean number of steps (y-axis) needed to find the goal is plotted per trial for two static, two dynamic and two inverse-dynamic agents (see text for explanation).

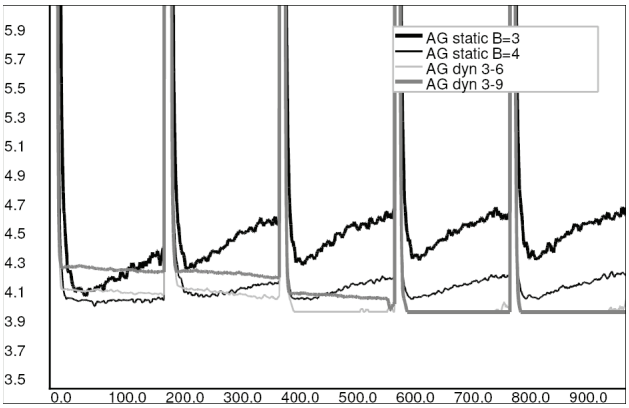

Figure 3.6. Alternating-Goal task; convergence plots of all learning phases (*n*=800), phases start at 0, 200, etc. 800. The mean number of steps (*y*-axis) needed to find the goal is plotted per trial for two static, two dynamic and two inverse-dynamic agents (see text for explanation).

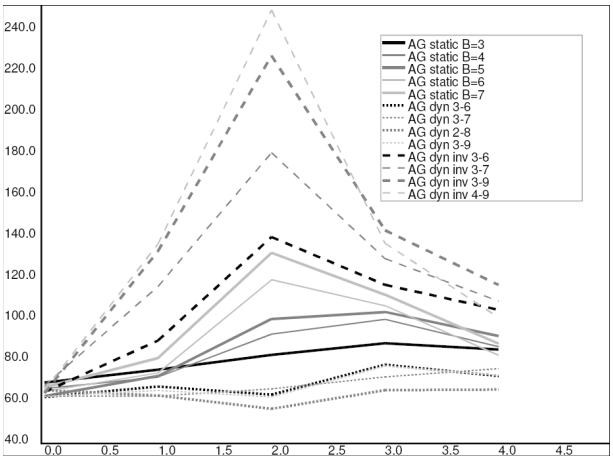

Figure 3.7. Alternating-Goal task; mean peaks of phases 0 to 4 (steps needed *at* respectively trail 0, 200, 400, 600 end 800) (*n*=800). Phase is on *x*-axis (only the integers); mean number of steps is on *y*-axis. The graph shows an overview for all agents of the mean number of steps needed to find the goal *at the goal switch*.

3.3.2 Experiment 2: Candy Task

Type-A agents have a considerable adaptation benefit compared to both control and type-B agents as shown by the following. In general, type-A agents have the same speed of finding the candy as exploiting agents (agents with a high static β), as shown by the learning curves of the complete task (Figure 3.8) and by the detailed learning curves of the start of the Candy task (Figure 3.10). In both figures the learning curves of β=6, and β=10 and dyn 2-8 overlap considerably. Interestingly, the quality of life curves show that in the beginning the QOL of the type-A agent quickly converges to the local optimum (candy, 0.25) comparable to that of the high β control agent (Figure 3.11, left "knee"). At the end of the task (later trials) the QOL of the type-A agent steadily increases towards the global optimum (food, +1.0; Figure 3.9). This shows that type-A affective feedback helps to first exploit a local optimum, while at a later stage explore for and exploit a global optimum. This is a major adaptation benefit resulting from type-A affective control of exploration. A playful way to think about this, is that the

agent "gets bored" with the local optimum and as a result starts to search for other things, thereby increasing the chance of finding the global optimum.

The control agent with $\beta = 4$ does converge to the global optimum just like the type-A agent (Figure 3.9). However, due to continuous high randomness in this agents action-selection mechanism this agent consistently needs more steps to get to that global optimum as compared to the type-A agent (Figure 3.8). Also due to this high randomness this agent does not learn the local optimum consistently enough to quickly exploit it (Figure 3.11). High static exploration (smaller βs) results in a major delay in arriving at the same level of QOL as compared to the larger βs and the type-A agent (compare "candy static 4" curve with "candy dyn 2-8" curve in Figure 3.11). The type-B agent does not perform well at converging or at quickly exploiting the local optimum (Figure 3.8, 3.9, 3.10, 3.11).

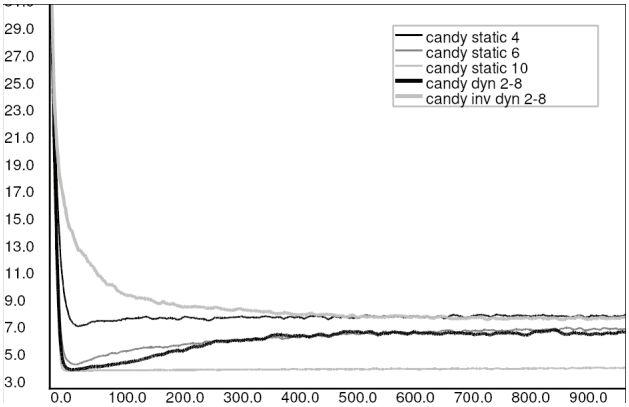

Figure 3.8. Candy task complete, mean learning curves (n=400). The mean number of steps (y-axis) needed to find the goal is plotted per trial for three static agents, one dynamic and one inverse-dynamic agent (see text for explanation).

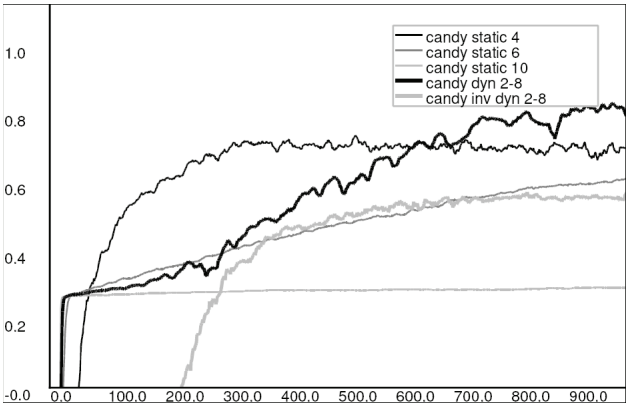

Figure 3.9. Candy task complete, mean Quality of Life curves (n=400). The mean QOL (y-axis) as it varies per trial is plotted for three static agents, one dynamic and one inverse-dynamic agent (see text for explanation).

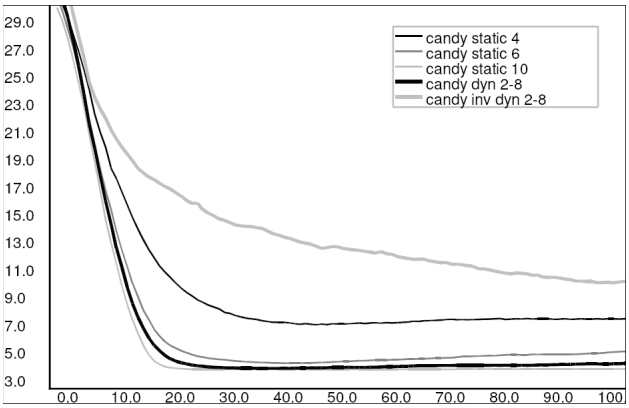

Figure 3.10. Candy task starts learning, mean learning curves (*n*=400). The mean number of steps (*y*-axis) needed to find the goal is plotted per trial for three static agents, one dynamic and one inverse-dynamic agent (see text for explanation).

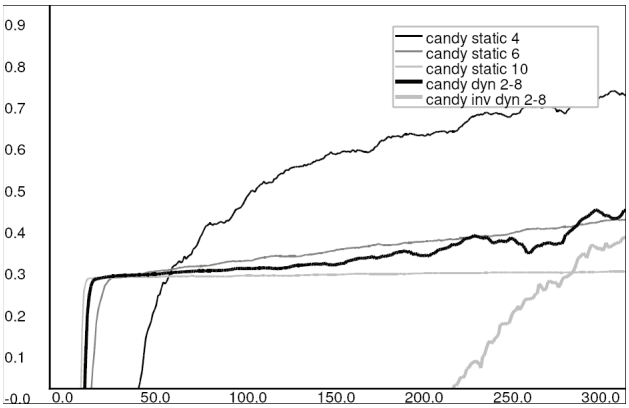

Figure 3.11. Candy task starts learning, mean Quality of Life curves (*n*=400). The mean QOL (*y*-axis) as it varies per trial is plotted for three static agents, one dynamic and one inverse-dynamic agent (see text for explanation).

3.4 General Discussion

Our results show that coupling positive affect to exploitation and negative affect to exploration provides two important benefits to learning and adaptation in the particular case of the grid worlds we have tested. Agents that use affect to control exploration in this way show significantly reduced task-switch cost *and* exploit a local optimum while being able to search for a global one.

3.4.1 Results Related to Other Learning Parameters

First, we briefly discuss the relation between learning and the proportion of local versus global optimum in the Candy task. If local and global optima are very similar, even a type-A agent cannot learn to prefer a global optimum, as the difference becomes very small. So, the candy and food reward have to be significantly different, such that the average β can exploit this difference once both options have been found. This has been confirmed in preliminary

experiments we conducted, and is quite plausible: “you don’t walk a long way for a little gain.”

Second, we discuss the relation between discount factor and learning. This relates to the previous; a small γ results in discarding rewards in the future and therefore the agent is more prone to fall for the nearer local optimum. So, γ should be set such that the agent is at least theoretically able to prefer a larger later reward for a smaller earlier one, which is also the reason why we incremented γ to 0.8 in the Candy task, as compared to 0.7 in the Alternating-Goal task.

3.4.2 Results Related to Psychological Findings

Our results illuminate several psychological findings. First and foremost, they show that to understand the relation between affect and learning, the *process* of affective influence on learning is important. Only by coupling affect to exploration versus exploitation were we able to show that both positive and negative affect are useful for learning, but at different phases in the learning process. Negative affect induces exploration in those phases that need it, while positive affect induces exploitation of the learned model when needed. This is an important result providing empirical evidence (albeit simulated) for the idea that both negative and positive affect can relate to faster learning (Craig et al., 2004). It also provides evidence for the claim that some aspects of negative emotions are useful mechanisms for adaptation (Hecker & Meiser, 2005). More specifically, negative affect can defocus attention and thereby favor less selective decision making (Hecker & Meiser, 2005) (in our study modeled as a more random choice of action). Our results show that the dynamic coupling of affect and decision-making can increase adaptive potential of an agent if (1) negative affect relates to less selective decision making and (2) positive affect relates to more selective decision making.

Our results seem incompatible with the results by Dreisbach and Goschke (2004). They (and others) find that positive affect is related to more flexible, more distractible behaviors. In short, they argue that positive affect decreases selectivity (by increasing flexibility and distractibility) while negative affect increases selectivity (by decreasing flexibility and decreasing distractibility). However, closer investigation of their empirical results allows for a plausible alternative interpretation that relates to normal conditioning (Reinforcement Learning) effects. We discuss this is in detail, as our alternative explanation potentially is relevant to many affect induction tasks that measure reaction time and that allow for subjects to get accustomed to the task while it is being performed.

Dreisbach and Goschke (2004) measure the difference between reaction time (RT) before a task switch and after a task switch. This difference is interpreted as switch cost. So, if a task takes 600 ms at trials before a change in the characteristics of a task and 700 ms after that change, than this difference (100 ms) is the switch cost. The experimental setup is as follows. During a set of trials, subjects have to perform a simple cognitive task proposed in the target color (e.g., red). At the same time they see a different instance of the same cognitive task in the distracter color (e.g., blue). The subject's task is to react only to the task in the target color. Half way, there is a task switch. Now, two situations are possible, perseverance and learned irrelevance. In the perseverance condition, the target task is presented in a new color (e.g., yellow) and the distracter task is presented in the old target color. The subject's challenge is to *not* continue solving the task in the old target color. In the learned irrelevance condition the target is presented in the old distracter color (blue), while the distracter task is presented in a new color (yellow). The challenge here is not to be hindered by the novel color yellow or be inhibited by the old distracter color blue that has become the target color.

The main thrust for Dreisbach and Goschke's conclusion that positive affect reduces perseveration (= continuation on an old goal) but increases flexibility (= potential to switch to a new cognitive set) is (1) the relative lack of switch cost in the perseveration condition and (2) the increase of switch cost in the learned irrelevance condition. They argue that this is a specific effect of affect on perseveration versus flexibility. We will now present an alternative explanation based on standard learning and conditioning effects.

Affect can be interpreted as an unattributed reinforcement signal. First, it is generally accepted that floating (objectless) positive and negative affect is a signal to the organism defining the general goodness versus badness of the situation (e.g., Gasper & Clore, 2002). Second, we have argued and shown experimentally that reinforcement and affect are strongly related. Third, affect is coupled to the dopamine system (Ashby et al., 1999)—a system that is also highly related to Reinforcement Learning, a point explicitly made by, and one that underlies Dreisbach and Goschke's (2004) approach.

Therefore affect induction can alternatively be understood as unconscious reinforcement of trials. So in, e.g., the study by Dreisbach & Goschke (2004) positive affect induction can be seen as conditioning upon a certain task, specifically as the trials are repeated many times before the task switch is introduced. This means that subjects actually learn differently when affectively induced as compared to control or non-affective situations. This is an important point underlying our alterative interpretation.

Consider the following. When positive affect is induced, the subject is actually reinforced to respond to the task presented in the target color red and *not* to respond to the task presented in the distracter color blue. After the switch to the perseveration condition, the new color yellow is introduced (and the subject is explicitly made aware of this change). Now, there are two tasks. A new, neutral—non-reinforced—colored task and an old positively-reinforced colored task.

Consider the switch to the learned irrelevance condition. Again the subject is first reinforced on the target color red, and the task switch introduces the new color yellow. However, the distracter is presented in yellow, while the new target is presented in the old distracter color blue. This means that in the first condition the subject learns to react to a new stimulus (yellow), while in the second it has to perform reversal learning (blue meant no action, but now it means action). Reversal learning is generally considered more difficult than learning new behavior. According to this explanation, in the perseveration condition one would expect slightly better learning of the post-switch condition due to the generic effect of positive reinforcement during learning. In the learned irrelevance condition one would expect a much worse learning of the post-switch condition due to unlearning (reversal learning). This is almost exactly what has been found, *if the results are combined with a generic negative influence of positive affect on RT*. First, all positive affect situations have slightly higher RTs than the control (and pre-) tests, reflecting a negative influence of positive affect on performance on this specific task. Second, the perseveration condition has lower post-switch cost in the positive affect situation compared to the control (and pre-) test, reflecting enhanced learning due to positive affect. Third, the learned irrelevance condition has a major increase in switch cost as compared to the control (and pre-) tests, reflecting difficulty unlearning the previous association between distracter color and irrelevance.

This alternative explanation is plausible, albeit speculative. The main message of this elaborate discussion is that many affect induction studies could be measuring confounded dependent variables. The measured total effect can be a combination of both a learning-related effect (conditioning) and a top-down executive control effect that is not specifically related to learning (working memory, etc.). This is particularly important as these studies are done to measure the second effect. If part of the total effect attributed to top-down influences is in fact due to bottom-up influences, it is highly important to control for the bottom-up effect. The results of our—quite unusual—bottom-up approach to model a phenomenon that is typically considered top-down, shows that reasonably simple, and arguably low-level effects *can* be responsible for part of the flexibility effect. An additional experimental problem arises when attempting to separate these two

effects, as both affect and reward seem to be mediated by the same dopamine system (Ashby et al., 1999). To summarize, our results cannot, at least not without further study, be considered as contrasting to results such as the ones discussed.

Current discussion on the Iowa Gambling Task (IGT) highly relates to our alternative explanation for the Dreisbach and Goschke study given here. The IGT (Bechara et al., 1997) measures the extent to which subjects learn to prefer to select cards from good decks versus bad decks. Good decks have many cards with low immediate monetary gain and some cards with low monetary loss. Bad decks have many cards with high immediate monetary gain but some cards with even higher loss. Overall, selecting cards from bad decks results in an average loss, while selecting cards from good decks results in an average gain. Subjects are unaware of the difference between decks and are asked to maximize gain by selecting cards from 4 decks (2 good, 2 bad).

In a sense, the IGT measures task-switching behavior. Up until the first bad card is selected from a bad card deck, these decks appear good, as they propose higher immediate monetary rewards than the good decks. After having selected the first bad card, subjects should re-evaluate (either consciously or unconsciously) their selection bias, ideally resulting in card-selection behavior directed at good decks. As subjects do not have any knowledge of the decks, we can easily interpret selecting the first bad card as a rule change that changes the current task. Prefrontal patients have difficulty learning to select cards from good decks instead of bad decks (Bechara et al., 1997). Alternatively, one could say that these patients are unable to switch to the new task of selecting from a good deck after having been reinforced to select from a bad deck. This interpretation suggests that prefrontal patients have difficulty switching to a new task in a Reinforcement Learning setting, which is quite plausible as the prefrontal cortex is often associated with executive control. Such control is needed for exactly this kind of task switches. This task-switch deficit might result from a lacking somatic marker signal (Damasio, 1994). However, in a recent review (Dunn, Dalgleish & Lawrence, 2006) it is argued that a reversal learning deficit can provide an alternative explanation. In a broad sense, this indicates that reversal learning is an important phenomenon to consider in all experiments that use (1) a learning task with potential involvement of reinforcement or affect, and (2) a, to the subject unknown, task switch due to a rule change. In a narrow sense, reversal learning is important in affect-induction cognitive-set switching experiments.

A comparison between the IGT and the Candy task is in place. The IGT has 4 decks of which 2 are good and 2 are bad. Every deck has a distribution of gain

and loss cards. In terms of Reinforcement Learning one could say that a subject needs exploration to build a model of the average gain of the decks and subsequently needs exploitation to continue selecting cards from the good decks. Three main issues are thus involved in learning the IGT: (1) build a model of the goodness of a deck, (2) vary between decks such that all decks are covered, and (3) exploit the knowledge gained.

The Candy task is different (and simpler). There are no changing rewards. There is a local and a global maximum. The agent has to learn, through exploration, that a global maximum exists, and then exploit this maximum. Exploration-exploitation is controlled by affect in our studies. Key difference between the Candy task and the IGT thus is that the rewards in the Candy task are deterministic, i.e., once the agent has found the reward, it knows that this is the correct reward for that location in the maze. In the Candy task only the second and third issues are important (exploration-exploitation). Since the varying rewards in the IGT are a key characteristic of that task, our Candy task cannot be considered analogous to the IGT. Future work includes measuring the behavior of agents that use affect to control exploration-exploitation in a simulated IGT.

Of course alternative explanations for our experimental results are possible. Our model for affect could, for example, be interpreted as a model for *flow* (Csikszentmihalyi, 1990). If reward is consistently better than expected, we are in a state of flow and therefore continue to do what we do (model-based decisions). If reward is consistently worse than expected, we are out of flow, and engage in more random, search-like behavior.

However, our model of affect does seem to have face-validity, specifically in the context of adaptation. If things go well, don't change. If things go bad, explore alternatives. This kind of underlying principle is quite plausible, but in stark contrast to the following: if positive affect indicates goodness, *we can afford to explore*, and if negative affect indicates badness, *we should be very selective regarding our behavior* (Dreisbach & Goschke, 2004). Which relation between affect and adaptation is right? Probably both, and the question is when and in what tasks? Only more elaborate process-oriented experimental and simulation studies will be able to show.

3.4.3 Results Related to Exploration and Exploitation in Machine Learning

The merit of using artificial affect as controller for exploration versus exploitation behavior has to be seen in light of adaptive agents in potentially changing environments. Such agents ideally decide autonomously when to explore versus

exploit. It is in this context that we propose affect as signal to guide the learning process.

In contrast, standard methods exist that are far better at optimizing a solution to an arbitrary and *static* credit assignment problem. These methods stem from, e.g., operations research. Consider, for example, learning the optimal solution to the Candy task. This is merely a question of exploring enough in the beginning, and then gradually decreasing the amount of exploration; a process called *simulated annealing*. Given enough exploration and a smooth transition from exploration to exploitation, any RL mechanism is able to learn the optimal solution.

For an adaptive agent in a changing world, a gradual decrease in exploration is not what is needed mainly for two reasons. First, consider an autonomous robot that has to decide where to go. If that robot is purely exploring, it might choose actions that are lethal to it. In a simulated environment this is no problem, however, in a real environment this is. Second, consider a changing environment. In this case the problem is not static, and credit assignment can thus never reflect *the* optimal solution; it always reflects the *current* optimal solution. If a change occurs, the agent has to solve two problems that do not need to be solved for static problems. These are (a) how to detect the change, and (b) how to move back to exploration (in contrast to gradually moving from exploration to exploitation).

The problem we address with affect as meta-learning signal is not that of finding an optimal solution given an arbitrary problem. It is the problem of guiding the learning process such that the agent can autonomously decide *when* and *how* to explore versus exploit.

3.5 Conclusion

We have introduced a computational method of studying the relation between affect and probabilistic learning. Based on experimental results with learning agents in simulated grid worlds, we conclude that, at least in the task we have experimented with, coupling positive affect to exploitation and negative affect to exploration has two important adaptation-related benefits: 1) It significantly reduces the agent's "goal-switch search peak" when the agent learns to adapt to a new goal. The agent finds this new goal faster. 2) Artificial affect facilitates convergence to a global instead of a local optimum, while permitting to exploit that local optimum. Our results illuminate the process underlying the relation between affect and learning, and, we argue, is thereby a valuable addition to the existing affect-cognition literature. The results provide evidence for the idea that negative affect is related to less selective decisions while positive affect is related

to more selective decisions. Further, our reinforcement-learning based analysis showed a potential problem with affect-induction techniques: the measured total effect of positive affect can be a combination of both a learning-related effect (conditioning) and a top-down executive control effect that is not specifically related to learning (working memory, etc.). However, as we have experimented with (only) two different types of worlds, our conclusions can not be generalized. More research is needed.

From a machine learning perspective, we have shown that in some cases artificial affect can be useful to guide exploration versus exploitation. However, more experiments should be done, specifically in different, and larger, worlds, using other RL models (for example, models that are able to cope with continuous environments).