University of Massachussetts, Amherst, USA
Institute of Cognitive Sciences and Technologies, CNR, Roma, Italy
barto@cs.umass.edu

Intrinsic Motivation and Reinforcement Learning

Andrew G. Barto

May 13, 2012

Abstract

Motivation is a key factor in human learning. We learn best when we are highly motivated to learn. Psychologists distinguish between extrinsically-motivated behavior, which is behavior undertaken to achieve some externally supplied reward, such as a prize, a high grade, or a high-paying job, and intrinsically-motivated behavior, which is behavior done for its own sake. Is an analogous distinction meaningful for machine learning systems? Can we say of a machine learning system that it is motivated to learn, and if so, is it possible to provide it with an analog of intrinsic motivation? Despite the fact that a formal distinction between extrinsic and intrinsic motivation is elusive, this chapter argues that the answer to both questions is assuredly "yes," and that the machine learning framework of reinforcement learning is particularly appropriate for bringing learning together with what in animals one would call motivation. Despite the common perception that a reinforcement learning agent's reward has to be extrinsic because the agent has a distinct input channel for reward, reinforcement learning provides a natural framework for incorporating principles of intrinsic motivation.

1 Introduction

Motivation refers to processes that influence the arousal, strength, and direction of behavior. "To be motivated means *to be moved* to do something" (Ryan and Deci 2000). Psychologists distinguish between *extrinsic motivation*, which means doing something because of some externally supplied reward, and *intrinsic motivation*, which refers to "doing something because it is inherently interesting or enjoyable" (Ryan and Deci 2000). Intrinsic motivation leads organisms to engage in exploration, play, and other behavior driven by curiosity in the absence of externally-supplied rewards.

This chapter focuses on how to frame concepts related to intrinsic motivation using the computational theory of reinforcement learning (RL) as it is studied by machine learning researchers (Sutton and Barto 1998). It is a common perception that the computational RL framework[1] can only deal with extrinsic

[1]The phrase *computational RL* is used here because this framework is not a theory of

motivation because an RL agent has a distinct input channel that delivers reward from its external environment. In contrast to this view, this chapter argues that this perception is a result of not fully appreciating the abstract nature of the RL framework, which is, in fact, particularly well suited for incorporating principles of intrinsic motivation. It further argues that incorporating computational analogs of intrinsic motivation into RL systems opens the door to a very fruitful avenue for the further development of machine learning systems.

RL is a very active area of machine learning, with considerable attention also being received from decision theory, operations research, and control engineering, where it has been called "Heuristic Dynamic Programming" (Werbos 1987) and "Neuro-Dynamic Programming" (Bertsekas and Tsitsiklis 1996). There is also growing interest in neuroscience because the behavior of some of the basic RL algorithms closely correspond to the activity of dopamine producing neurons in the brain, as described elsewhere in this volume. RL algorithms address the problem of how a behaving agent can learn to approximate an optimal behavioral strategy, usually called a *policy*, while interacting directly with its environment. Viewed in the terms of control engineering, RL consists of methods for approximating closed-loop solutions to optimal control problems while the controller is interacting with the system being controlled. This engineering problem's optimality criterion, or objective function, is analogous to the machinery that delivers primary reward signals to an animal's nervous system. The approximate solution to the optimal control problem corresponds to an animal's skill in performing the control task. A brief introduction to RL is provided in Section 2 below.

Providing artificial learning systems with analogs of intrinsic motivation is not new. Lenat's AM system (Lenat 1976), for example, included an analog of intrinsic motivation that directed search using heuristic definitions of "interestingness." Schmidhuber (1991a, 1991b, 1997, 1999) introduced methods for implementing forms of curiosity using RL algorithms, and Sutton's (1991) "exploration bonus" is a form of intrinsic reward. The author's efforts on this topic began when he and colleagues realized that some new developments in RL could be used to make intrinsically-motivated behavior a key factor in producing more capable learning systems. This approach, introduced by Barto et al. (2004) and Singh et al. (2005), combines intrinsic motivation with methods for temporal abstraction introduced by Sutton et al. (1999). The reader should consult Barto et al. (2004) and Singh et al. (2005) for more details on this approach.

Not all aspects of motivation involve learning—an animal can be motivated by innate mechanisms that trigger fixed behavior patterns, as the ethologists have emphasized—but what many researchers mean by motivated behavior is behavior that involves the assessment of the consequences of behavior through learned expectations (e.g., Epstein 1982, McFarland and Bösser 1993, Savage 2000). Motivational theories therefore tend to be intimately linked to theories of learning and decision making. Because RL addresses how predictive values

biological RL despite what it borrows from, and suggests about, biological RL. Throughout this chapter RL refers to computational RL.

can be learned and used to direct behavior, RL is naturally relevant to the study of motivation.[2]

The starting point for addressing intrinsic motivation using the RL framework is the idea that learning and behavior generation processes "don't care" if the reward signals are intrinsic or extrinsic (whatever that distinction may actually mean!); the same processes can be used for both. Schmidhuber (1991a) put this succinctly for the case of curiosity and boredom: "The important point is: The same complex mechanism which is used for 'normal' goal-directed learning is used for implementing curiosity and boredom. There is no need for devising a separate system ..." As we shall see in what follows, this idea needs clarification, but it underlies all of the approaches to using RL to obtain analogs of intrinsic motivation: it becomes simply a matter of defining specific mechanisms for generating reward signals.

Although this approach is attractive in its simplicity and accords well with prevalent—though controversial—views on the pervasive influence of brain reward systems on behavior (e.g., Linden 2011), other theoretical principles—not discussed here—have been proposed that can account for aspects of intrinsic motivation, e.g., Andry et al. (2004), Baranes and Oudeyer (2010), Friston et al. (2010), Hesse et al. (2009). Further, contemporary psychology and neuroscience indicate that the nature of reward signals is only one component of the complex processes involved in motivation (Daw and Shohamy 2008). Despite these qualifications, restricting attention to the nature of reward signals within the RL framework illuminates significant issues for the development of computational analogs of motivational processes.

Several important topics relevant to motivation and intrinsic motivation are beyond this chapter's scope. There has been a great increase in interest in affect and emotion in constructing intelligent systems (e.g., Picard 1997, Trappl et al. 1997). Motivation and emotion are intimately linked, but this chapter does not address computational theories of emotion because it would take us too far from the main focus. Also not discussed are social aspects of motivation, which involve imitation, adult scaffolding, and collective intentionality, all of which play important roles in development (e.g., Breazeal et al. 2004, Thomaz and Breazeal 2006, Thomaz et al. 2006).

This chapter also does not attempt to review the full range of research on motivational systems for artificial agents, to which the reader is referred to the extensive review by Savage (2000). Even research that explicitly aims to combine intrinsic motivation with RL has grown so large that a thorough review is beyond the scope of this chapter. The reader is referred to Oudeyer and Kaplan (2007) and Oudeyer et al. (2007) for a review and perspective on this research.

The concept of motivation in experimental and psychoanalytic psychology as well as in ethology has a very long and complex history that is discussed in many books, for example, those by Arkes-Garske (1982), Beck (1983), Cofer

[2]RL certainly does not exclude analogs of innate behavioral patterns in artificial agents. The success of many systems using RL methods depends on the careful definition of innate behaviors, as in the work of Grupen and colleagues described elsewhere in this volume.

and Appley (1964), Deci and Ryan (1985), Klein (1982), Petri (1981), and Toates (1986). This chapter only touches the surface of this extensive topic, with the goal of giving a minimal account of issues that seem most relevant to computational interpretations of intrinsic motivation and to describe some of the old theories to which the recent computational approaches seem most closely related.

This chapter begins with a brief introduction to the conventional view of RL, which is followed by two sections that provide some historical background on studies of motivation and intrinsic motivation. These are followed by two sections that respectively relate RL to motivation in general and intrinsic motivation in particular. Discussion of what an evolutionary perspective suggests about intrinsic motivation is next, followed by a brief summary, discussion of prospects, and finally some concluding comments.

2 Reinforcement Learning

RL refers to improving performance through trial–and–error experience. The term reinforcement comes from studies of animal learning in experimental psychology, where it refers to the occurrence of an event, in the proper relation to a response, that tends to increase the probability that the response will occur again in the same situation (Kimble 1961). Although the specific term "reinforcement learning" is not used by psychologists, it has been widely adopted by theorists in artificial intelligence and engineering to refer to a class of learning tasks and algorithms. Early uses of this term were by Minsky (1961), Waltz and Fu (1965), and Mendel and McLaren (1970) in describing approaches to learning motivated in part by animal learning studies. The simplest RL algorithms are based on the commonsense idea that if an action is followed by a satisfactory state of affairs, or an improvement in the state of affairs, then the tendency to produce that action is strengthened, i.e., reinforced, following Thorndike's (1911) "Law of Effect."

The usual view of an RL agent interacting with its environment is shown in Figure 1. The agent generates actions in the context of sensed states of this environment, and its actions influence how the environment's states change over time. This interaction is typically viewed as happening in discrete time steps, $t = 1, 2, 3, \ldots$, that do not represent the passage of any specific amount of real time. The environment contains a 'critic' that provides the agent at each time step with an evaluation (a numerical score) of its ongoing behavior.[3] The critic maps environment states (or possibly state-action pairs or even state-action-next-state triples) to numerical reward signals. The agent learns to improve its skill in controlling the environment in the sense of learning how to cause larger magnitude reward signals to be delivered from its environment over time. The information the critic conveys to the agent corresponds to information

[3]The term critic is used, and not 'teacher', because in machine learning a teacher provides more informative instructional information, such as directly telling the agent what its actions *should have been* instead of merely scoring them.

about what psychologists call *primary reward*, generally meaning reward that encourages behavior directly related to survival and reproductive success, such as eating, drinking, escaping, etc. The mapping from states to reward signals implemented by the critic is called a *reward function*. In RL the reward function is an essential component in specifying the problem the agent must learn to solve.

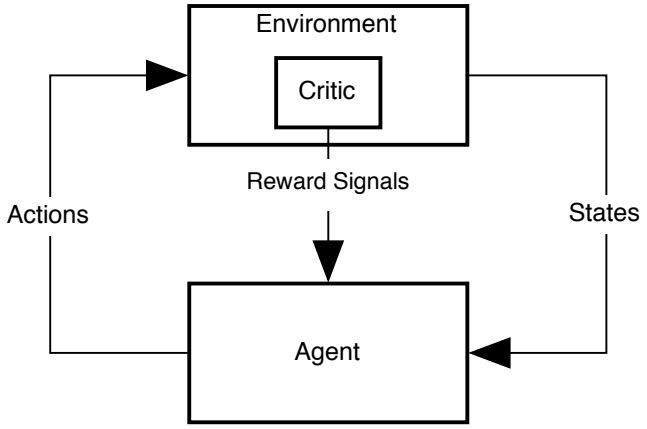

Figure 1. *Agent-Environment Interaction in RL. Primary reward signals are supplied to the agent from a "critic" in its environment. Adapted from Barto et al. (2004).*

The agent's specific objective is to act at each moment of time so as to maximize a measure of the total quantity of reward it expects to receive over the future. This measure can be a simple sum of the reward signals it expects to receive over the future, or more frequently, a discounted sum in which later reward signals are weighted less than earlier ones. The value of this measure at any time is the agent's *expected return*. Because the agent's actions influence how the environment's state changes over time, maximizing expected return requires the agent to exert control over the evolution of its environment's states. This can be very challenging. For example, the agent might have to sacrifice short-term reward in order to achieve more reward over the long-term. The simplest RL agent's attempt to achieve this objective by adjusting a *policy*, which is a rule that associates actions to observed environment states. A policy corresponds to a stimulus-response (S-R) rule of animal learning theory. But RL is not restricted to simple S-R agents: more complicated RL agents learn models of their environments that they can use to make plans about how to act appropriately.

Note that since return is a summation, a reward signal equal to zero does not contribute to it. Thus, despite the fact that the critic provides a signal at

every moment of time, a signal of zero means "no reward." Many problems are characterized by reward signals for which non-zero values are relatively rare, occurring, for example, only after the completion of a long sequence of actions. This is called the *problem of delayed rewards*, and much of RL is devoted to making learning efficient under these conditions.

The approach that has received the most attention focuses on RL agents that learn to predict return and then use these predictions to evaluate actions and to update their policies instead of using the primary reward signal itself. For example, in one class of methods, called *adaptive critic* methods (Barto et al. 1983), the agent contains a prediction component—an adaptive critic—that learns to predict return. An action that improves the likelihood of obtaining high return, as predicted by the adaptive critic, is reinforced. An increase in the prediction of return, then, acts as a reward itself. With these methods learning does not have to wait until a final goal is achieved.[4] This predictive ability of an adaptive critic mimics the phenomenon of *secondary*, or *conditioned*, reinforcement observed in animal learning (Mackintosh 1983). A secondary reinforcer is a stimulus that has become a reinforcer by virtue of being repeatedly paired in an appropriate temporal relationship with either a primary reinforcer or with another secondary reinforcer. In other words, a secondary reinforcer is a stimulus that has acquired, through a learning process, the ability to act as reinforcer itself.

Before going further it is critical to comment on how this abstract RL formulation relates our view of an animal or a robot. An RL agent should not be thought of as an entire animal or robot. It should instead be thought of as the component *within* an animal or robot that handles reward-based learning. Thus the box labeled "Environment" in Figure 1 represents not only what is in the animal or robot's external world, but also what is external to the reward-based learning component *while still being inside the animal or robot*. In particular, the critic in Figure 1 should be thought of as part of an animal's nervous system and not as something in the animal's external world. Similarly, an RL agent's "actions" are not necessarily like an animal or robot's overt motor actions; they can also be actions that affect the agent's internal environment, such as the secretion of a hormone or the adjustment of a processing parameter.

It is also important to note that although the critic's signal at any time step is usually called a "reward" in the RL literature, it is better to call it a "reward signal" as it is labeled in Figure 1. The reason for this is that psychologists and neuroscientists distinguish between rewards and reward signals. Schultz (2007a, 2007b), for example, writes: "Rewards are objects or events that make us come back for more" whereas reward signals are produced by reward neurons in the brain. It is much more appropriate to think of the critic's signal as analogous to the output of a brain reward system than as an object or event in the animal's external world. These observations are important for understanding how the RL framework accommodates intrinsic reward signals and will be returned to

[4]It is important to note that the adaptive critic of these methods is *inside* the RL agent, while the different critic shown in Figure 1—that provides the primary reward signal—is in the RL agent's environment.

in Section 6 below.

Despite the fact that its roots are in theories of animal learning, RL is—with some exceptions—a collection of computational tools for use in artificial systems rather than a collection of animal behavioral models. A wide range of facts about animal motivation are not usefully captured by the current RL framework. Dayan (2001), for example, correctly comments as follows:

> "Reinforcement learning (RL) bears a tortuous relationship with historical and contemporary ideas in classical and instrumental conditioning. Although RL sheds important light in some murky areas, it has paid less attention to research concerning *motivation* of stimulus-response (SR) links."

A major reason for this neglect is that the mathematical framework of RL, as it is conventionally formulated (Sutton and Barto 1998), takes the existence of a reward signal as a given: the theory is not concerned with processes that generate reward signals. All that a well-posed RL problem requires is the specification of some (bounded) real-valued function from states to reward signals (or, in some cases, from state-action pairs, or from state-action-next-state triples, to reward signals). This not only sidesteps the entire subject of utility theory, which relates scalar measures to agent preferences, it also sidesteps many of the issues relevant to what (for an animal) would be called motivation.

Instead of being a shortcoming of the conventional RL framework, however, this level of abstraction has been a distinct advantage. It has allowed the theory of RL to progress in the absence of specific assumptions about how reward signals are generated in special cases. As a result, RL has been useful for a great many different types of problems, and it readily lends itself to being incorporated into a wide variety of comprehensive architectures for autonomous agents, in each of which different assumptions are made about how reward signals are generated. The abstract nature of RL is perhaps a major reason that it has been able to shed important light, as Dayan remarked, on some murky areas of biological data. Luckily, an account of intrinsic motivation in RL terms can be produced with only minor reduction in the framework's level of abstraction by introducing some assumptions about the nature of an RL agent's environment and reward signals. This is taken up in Section 6 below.

3 Motivation

Describing what the "hypothetical man on the street" means when asking why someone has behaved in a particular way, Cofer and Appley (1964) list three categories of factors: (1) irresistible external influences, (2) an internal urge, want, need, drive, plan, etc. or (3) an external object or situation acting as a goal, or incentive. The first of these exert their influence largely independently of the internal state of the organism as, for example, a reflexive withdrawal from a painful stimulus. The second two, in contrast, involve hypothesized internal states regarded as being necessary to explain the behavior. Incentive objects

are external, but are endowed with their behavior-controlling ability through the assignment to them of a state-dependent value by the organism. Motivational explanations of the strength and direction of behavior invoke an organism's internal state.

A clear example of the influence of internal motivational state on behavior is an experiment by Mollenauer (1971) as described by Dickinson and Balleine (2002). Rats were trained to run along an alleyway to obtain food. Rats in one group were trained while hungry, being food deprived before each training session, while rats in another group were nondeprived. The hungry rats consistently ran faster than did the sated rats. It might simply be that when rats are trained while they are hungry, they tend to run faster when the results of learning are tested. But the second part of Mollenauer's experiment showed that a shift in deprivation state had an immediate effect on the rat's performance. Rats in a third group were trained while hungry but tested when nondeprived. These rats immediately ran slower after this motivational shift. Instead of having to experience reduced reward for eating in the nondeprived state, their nondeprived state somehow exerted a direct and immediate influence on behavior. The kind of rapid behavioral change illustrated in this experiment and many others required theorists to postulate the existence of multiple internal motivational states. This experiment also illustrates the view taken by psychologists studying animal learning about how motivation and learning are intimately linked. Motivational factors can influence learning through their control over the effectiveness of reward and their control over how the results of learning are expressed in behavior.

The starting point for including motivational factors in the RL framework is to be clear about what we mean by an "internal state." In an extensive review of motivation for artificial agents, Savage (2000) focused on an "interactive view of motivation," attributed to Bindra (1978) and Toates (1986), that explains motivation in terms of a *central motive state* that depends on the interaction of an internal state and an external incentive factor:

$$\text{central motive state} = (\text{internal state}) \times (\text{incentive factor})$$

In Bindra's (1978) account, a central motive state arises through the interaction of an internal "organismic state" (such as arousal level, blood-sugar level, cellular dehydration, estrus related hormonal levels) and features of an incentive object (such as features indicating the palatability of a food object).

The elaboration of the RL framework in Section 6 below roughly follows the interactive approach by factoring the RL problem's state into two components: a component internal to the animal or robot, and a component external to the animal or robot. This means that the RL problem's state at any time t is represented as a vector $s_t = (s_t^i, s_t^e)$, where s_t^i and s_t^e are respectively the internal (cf. "organismic") and external state components (each of which can itself be a vector of many descriptive feature values). The nature of the dependency of reward signals on the internal dimensions is of particular importance for including intrinsic motivational factors in the RL framework.

There are certainly many grounds for disputing this view of motivation, but at a commonsense level it should be clear what is intended. If an organism is active in the sense of not being driven totally by environmental stimuli—a view that by now must be universal—then the organism must not implement a memoryless mapping from stimuli to responses, that is, there must be more than one internal state. Going further, McFarland and Bösser (1993) argue that for motivational descriptions of behavior to be meaningful, the agent has to have some degree of autonomy, that is, it must be capable of self-control, by which they mean that changes in behavior are the result of explicit decision processes that weigh behavioral alternatives. Thus, it would not be useful to talk about the behavior of a clockwork automaton in motivational terms even though it may have many internal states.

Among the influential theories of motivation in psychology are the drive theories of Hull (1943, 1951, 1952). According to Hull, all behavior is motivated either by an organism's survival and reproductive needs giving rise to primary drives (such as hunger, thirst, sex, and the avoidance of pain) or by derivative drives that have acquired their motivational significance through learning. Primary drives are the result of physiological deficits—"tissue needs"—, and they energize behavior whose result is to reduce the deficit. A key additional feature of Hull's theories is that a need reduction, and hence a drive reduction, acts as a primary reward for learning: behavior that reduces a primary drive is reinforced. Additionally, through the process of secondary reinforcement in which a neutral stimulus is paired with a primary reward, the formerly neutral stimulus acquires the reinforcing power of the primary reward. In this way, stimuli that predict primary reward, i.e., a reduction in a primary drive, become rewarding themselves. Thus, according to Hull, all behavior is energized and directed by its relevance to primal drives, either directly or as the result of learning through secondary reinforcement.

Hull's theories follow principles adapted from the concept of physiological homeostasis, the term introduced by Cannon (1932) to describe the condition in which bodily conditions are maintained in approximate equilibrium despite external perturbations. Homeostasis is maintained by processes that trigger compensatory reactions when the value of a critical physiological variable departs from the range required to keep the animal alive. This negative feedback mechanism maintains these values within required bounds. Many other theories of motivation also incorporate, in one form or another, the idea of behavior being generated to counteract disturbances to an equilibrium condition.

Although many of their elements have not been supported by experimental data, this and related theories continue to influence current thinking about motivation. They have been especially influential in the design of motivational systems for artificial agents, as discussed in Savage's review of artificial motivational systems (Savage 2000). Hull's idea that reward is generated by drive reduction is commonly used to connect RL to a motivational system. Often this mechanism consists of monitoring a collection of important variables, such as power or fuel level, temperature, etc., and triggering appropriate behavior when certain thresholds are reached. Drive reduction is directly translated into

a reward signal for some type of RL algorithm.

Among other motivational theories are those based on the everyday experience that we engage in activities because we enjoy doing them: we seek pleasurable experiences and avoid unpleasant ones. These hedonic theories of motivation hold that it is necessary to refer to affective mental states to explain behavior, such as a "feeling" of pleasantness or unpleasantness. Hedonic theories are supported by many observations about food preferences which suggest that "palatability" might offer a more parsimonious account of food preferences than tissue needs (Young 1966). Animals will enthusiastically eat food that has no apparent positive influence on tissue needs; characteristics of food such as temperature and texture influence how much is eaten; animals that are not hungry still have preferences for different foods; animals have taste preferences from early infancy (Cofer and Appley 1964). In addition, non-deprived animals will work enthusiastically for electrical brain stimulation (Olds and Milner 1954).

Although it is clear that biologically-primal needs have motivational significance, facts such as these showed that factors other than primary biological needs exert strong motivational effects and that these factors do not derive their motivational potency as a result of learning processes involving secondary reinforcement.

4 Intrinsic Motivation

In addition to observations about animal food preferences and responses to electrical brain stimulation, other observations showed that something important was missing from drive reduction theories of motivation. Under certain conditions, for example, hungry rats would rather explore unfamiliar spaces than eat; they will endure the pain of crossing electrified grids to explore novel spaces; monkeys will bar-press for a chance to look out of a window. Moreover, the opportunity to explore can be used to reinforce other behavior. Deci and Ryan (1985) chronicle these and a collection of similar findings under the heading of *intrinsic motivation*.[5]

The role of intrinsically motivated behavior in both children and adults is commonly noted as, for example, in this quotation:

> The human organism is inherently active, and there is perhaps no place where this is more evident than in little children. They pick things up, shake them, smell them, taste them, throw them across the room, and keep asking, "What is this?" They are unendingly curious, and they want to see the effects of their actions. Children are intrinsically motivated to learn, to undertake challenges, and to solve problems. Adults are also intrinsically motivated to do a variety of things. They spend large amounts of time painting

[5] Deci and Ryan (1985) mention that the term intrinsic motivation was first used by Harlow in a 1950 study (Harlow 1950) showing that rhesus monkeys will spontaneously manipulate objects and work for hours to solve complicated mechanical puzzles without any explicit rewards.

> pictures, building furniture, playing sports, whittling wood, climbing mountains, and doing countless other things for which there are not obvious or appreciable external rewards. The rewards are inherent in the activity, and even though there may be secondary gains, the primary motivators are the spontaneous, internal experiences that accompany the behavior. (p. 11, Deci and Ryan 1985)

Why did most psychologists reject the view that exploration, manipulation, and other curiosity-related behaviors derived their motivational potency only through secondary reinforcement, as would be required by a theory like Hull's? There are clear experimental results showing that such behavior is motivationally energizing and rewarding on its own and not because it predicts the satisfaction of a primary biological need. Influential papers by White (1959) and Berlyne (**?**) marshaled abundant experimental evidence to argue that the intrinsic reward produced by a variety of behaviors involving curiosity and play are as primary as that produced by the conventional biologically relevant behaviors. Children spontaneously explore very soon after birth, so there is little opportunity for them to experience the extensive pairing of this behavior with the reduction of a biologically primary drive that would be required to account for their eagerness to explore. In addition, experimental results show that the opportunity to explore retains its energizing effect without needing to be repaired with a primary reward, whereas a secondary reward will extinguish, that is, will lose its reinforcing quality, unless often re-paired with the primary reward it predicts.

Berlyne summarized the situation as follows:

> As knowledge accumulated about the conditions that govern exploratory behavior and about how quickly it appears after birth, it seemed less and less likely that this behavior could be a derivative of hunger, thirst, sexual appetite, pain, fear of pain, and the like, or that stimuli sought through exploration are welcomed because they have previously accompanied satisfaction of these drives. (p. 26, **?**)

Note that the issue was not whether exploration, manipulation, and other curiosity-related behavior are important for an animal's survival and reproductive success. Clearly they are if deployed in the right way. Appropriately cautious exploration, for example, clearly contributes to survival and reproductive success because it can enable efficient foraging, successful escape, and increased opportunities for mating. The issue was whether these behaviors have motivational valence because previously *in the animal's lifetime* they predicted decreases in biologically-primary drives, or whether this valence is built-in by the evolutionary process. Section 7 looks more closely at the utility of intrinsic motivation from an evolutionary perspective.

Researchers took a variety of approaches in revising homeostatic drive reduction theories in light of findings like those described above. The simplest approach was to expand the list of primary drives by adding drives such as a curiosity drive, exploration drive, manipulation drive, etc., to the standard

list of drives. Postulating primary "needs" for these behaviors, on par with needs for food, drink, and sex, marked a break from the standard view while retaining the orthodox drive reduction principle. For example, an experiment by Harlow, Harlow, and Meyer (1950) showed that rhesus monkeys would learn how to unerringly unlatch a complex mechanical puzzle through many hours of manipulation without any contingent rewards such as food or water. They postulated a "strong and persistent manipulation drive" to explain how this was possible in the absence of extrinsic reward. Other experiments showed that giving an animal the opportunity to run in an activity wheel could act as reward for learning, suggesting that there is an "activity drive."

Postulating drives like these was in the tradition of earlier theorists who advanced broader hypotheses. For example, in a treatise on play, Groos (1901) proposed a motivational principle that we recognize as a major component of Piaget's (1952) theory of child development:

> The primitive impulse to extend the sphere of their power as far as possible leads men to the conquest and control of objects lying around them ... We demand a knowledge of effects, and to be ourselves the producers of effects. (p. 95, Groos 1901)

Similarly, Hendrick (1942) proposed an "instinct to master" by which an animal has "an inborn drive to do and to learn how to do."

In a 1959 paper that has been called "one of the most influential papers on motivation ever published" (Arkes and Garske 1982), Robert White (1959) argued that lengthening the list of primary drives in this way would require such fundamental changes to the drive concept as to leave it unrecognizable. Drives for exploration, manipulation, and activity do not involve "tissue needs"; they are not terminated by an explicit consummatory climax but rather tend to decrease gradually; and reinforcement can result from the increase in such a drive rather than a decrease: for example, when an exploring animal seeks out novelty rather than avoids it. If decreasing exploratory drive corresponds to boredom, one does not normally think of boredom as a reinforcer for exploration.

White proposed that instead of extending the list of the standard drives, it would be better to emphasize the similarity of urges toward exploration manipulation, and activity, and how they differ from the homeostatic drives. He proposed bringing them together under the general heading of *competence*, by which he meant *effective interaction with the environment*, and to speak of a general *effectance motivation* to refer to "an intrinsic need to deal with the environment." Like other critics of homeostatic theories, White did not argue that such theories are completely wrong, only that they are incomplete. With respect to effectance motivation, for example, he wrote that

> "... the effectance urge represents what the neuromuscular system wants to do when it is otherwise unoccupied or is gently stimulated by the environment. ... [it] is persistent in the sense that it regularly occupies the spare waking time between episodes of homeostatic crisis." (p. 321, White 1959)

The psychology literature is less helpful in specifying the concrete properties of experience that incite intrinsically motivated behavior, although there have been many suggestions. White (1959) suggested that Hebb (1949) may have been right in concluding that "difference-in-sameness" is the key to interest, meaning that along with many familiar features, a situation that is interesting also has novel ones, indicating that there is still more learning to be done. Berlyne (1954, 1960, 1971) probably had the most to say on these issues, suggesting that the factors underlying intrinsic motivational effects involve novelty (with a distinction between relative and absolute novelty, and between short-term and long-term novelty), or surprise and incongruity (when expectations or hypotheses are not vindicated by experience), or complexity (depending on the number and similarity of elements in a situation).

Uniting these cases, Berlyne hypothesized a notion of conflict created when a situation incites multiple processes that do not agree with one another. He also hypothesized that moderate levels of novelty (or more generally, arousal potential) have the highest hedonic value because the rewarding effect of novelty is overtaken by an aversive effect as novelty increases (as expressed by the "Wundt Curve", p. 89, of Berlyne 1971). This is consistent with many other views holding that situations intermediate between complete familiarity (boredom) and complete unfamiliarity (confusion) have the most hedonic value: the maximal effect of novelty being elicited by "... a stimulus that is rather like something well known but just distinct enough to be 'interesting.' " (Berlyne 1960). Many of these hypotheses fall under the heading of Optimal Level Theories, which we describe in more detail below.

The role of surprise in intrinsic motivation that Berlyne and others have suggested requires some discussion in order to prevent a common misunderstanding. Learning from surprise in the form of mismatch between expectations and actuality is built into many learning theories and machine learning algorithms. The Rescorla-Wagner (1972) model of classical conditioning and the many error-correction learning rules studied under the heading of *supervised learning* by machine learning researchers, such as perceptrons and error back-propagation neural networks, adjust parameters on the basis of discrepancies between expected and experienced input. Tolman's (1932) theory of latent learning, for example, postulated that animals are essentially *always learning* cognitive maps that incorporate confirmed expectancies about their environments. But according to Tolman, this learning is unmotivated; it does not depend on reinforcing stimuli or motivational state. This is different from Berlyne's view that surprise engages an animal's motivational systems. Therefore it is necessary to distinguish between learning from surprise as it appears in supervised learning algorithms and the idea that surprise engages motivational systems.

4.1 Optimal Level Theories

To provide alternatives to homeostatic drive reduction theories, and to avoid postulating additional drives for exploration, manipulation, etc., motivational theorists proposed a number of influential theories characterized as optimal

level theories. This section describes one of these at some length because it suggests useful principles anticipating several recent computational theories. This account is drawn largely from Arkes and Garske (1982).

Dember, Earl, and Paradise (1957) conducted an experiment involving an animal's preferences for stimuli of differing levels of complexity. They placed rats in a figure-eight runway having walls with vertical black and white stripes on one loop and horizontal black and white stripes on the other. To the moving rat, the horizontal stripes provided a roughly constant visual stimulus, whereas the vertical stripes provided a more complex time-varying stimulus pattern. With one rat in the runway at a time, they recorded the amount of time each rat spent in each loop. They found that a rat that initially preferred the loop with the horizontal stripes would later spend the preponderance of its time in the loop with the vertical stripes. Rats that initially preferred the vertical-striped loop rarely shifted preference to the horizontal-striped loop. In a another experiment, the horizontal stripes provided the more complex stimulus (compared to plain white or plain black walls), and the rats shifted preference to the horizontal stripes, thus ruling out the possibility that the behavior was due to some peculiarity of horizontal and vertical stripes.

Dember et al. (1957) proposed a theory (elaborated in Dember and Earl 1957) to explain this behavior. It is based on two key ideas. The first is that animals get used to a certain level of environmental complexity, and if they continue too long with stimuli of that complexity, they will become bored since they had already learned about stimuli of that complexity. A slightly more complex stimulus, on the other hand, will be interesting to them and will arouse curiosity, while an extremely more complex stimulus will be confusing, or even frightening. So an animal will maximally prefer stimuli that are moderately more complex than what they are used to. Dember and Earl used the term *pacer*, presumably from horse racing, to refer to the level of stimulus complexity that is maximally preferred. The second idea, which is common to other optimal level theories, is that as a result of experience with a stimulus, the stimulus becomes simpler to the animal. As Dember and Earl (1957) state, this is due to "the ability of stimuli to increase the psychological complexity of the individual who perceives them." Consequently, an animal's experience with a preferred stimulus situation causes their preferences to shift toward situations of moderately increased complexity: experience with a pacer causes the pacer to shift toward increased complexity. This generates a motivational force causing the animal to constantly seek stimuli of increasing complexity.

Berlyne's (1954, 1960, 1971) ideas, mentioned above, also fall under the heading of optimal level theory, and much of his work focused on trying to determine the stimulus properties that underlie an animal's preferences, as mentioned above. He regarded different properties, such as novelty, incongruity, and surprisingness, as contributing to the arousal potential of a stimulus, and that an animal will prefer some intermediate level of arousal potential.

Optimal level theories have been very influential, with applications in child development, where both excessive and deficient amounts of stimulation may be detrimental to cognitive growth, in architecture, city planning, esthetics,

economics, and music (Arkes and Garske 1982). The recent computational theory of Schmidhuber (2009) that places information-theoretic compression at the base of intrinsic motivation is a modern descendant of optimal level theories.

4.2 Intrinsic Motivation and Competence

In his classic paper, White (1959) argued that intrinsically motivated behavior is essential for an organism to gain the *competence* necessary for autonomy, where by autonomy he meant the extent to which an organism is able to bring its environment under its control, to achieve mastery over its environment. Through intrinsically motivated activity, an organism is able to learn much of what is possible in an environment and how to realize these possibilities. A system that is competent in this sense has *broad set of reusable skills* for controlling its environment; it is able to interact effectively with its environment toward a wide variety of ends. The activity through which these broad skills are learned is motivated by an intrinsic reward system that favors the development of broad competence rather than being directed to more specific externally-directed goals.

White's view of competence greatly influenced this author's thinking and that of his colleagues and students about the utility of analogs of intrinsic motivation for RL systems. Being competent in an environment by having a broad set of reusable skills enables an agent to efficiently learn how to solve a wide range of specific problems as they arise while it engages with that environment. Although the acquisition of competence is not driven by specific problems, this competence is routinely enlisted to solve many different specific problems over the agent's lifetime. The skills making up general competence act as the "building blocks" out of which an agent can form solutions to specific problems. Instead of facing each new challenge by trying to create a solution out of low-level primitives, it can focus on combining and adjusting its higher-level skills. This greatly increases the efficiency of learning to solve new problems, and it is a major reason that the relatively long developmental period of humans serves us so well.

By combining this view of competence with the theory and algorithms of hierarchical RL (Barto and Mahadevan 2003), the author and colleagues have taken some small steps toward developing artificial agents with this kind of competence, calling the approach *intrinsically motivated RL* (Barto et al. 2004, Singh et al. 2005). Evaluating the performance of these agents requires taking into account performance over ensembles of tasks instead of just single tasks. Intrinsically motivated RL therefore addresses the well-known shortcoming of many current machine learning systems—including RL systems—that they typically apply to single, isolated problems and do not yet allow systems to cope flexibly with new problems as they arise over extended periods of time. This brings the study of intrinsically motivated RL together with what cognitive scientists, roboticists, and machine learning researchers call "autonomous mental development" (Weng et al. 2001), "epigenetic robotics" (Prince et al. 2001), or "developmental robotics" (Lungarella et al. 2003), approaches aiming to develop analogs of the developmental processes that prepare animals over long

time periods for a wide range of later challenges.

This competence-based view of intrinsically motivated RL contrasts with the view most prominently put forward in the work of Schmidhuber (2009) that intrinsic motivation's sole purpose is to facilitate learning a world model in order to make accurate predictions. The competence view, in contrast, emphasizes the utility of learning skills that allow effective environmental control. Of course, learning such skills can benefit from the learning of models, and machinery for doing so is built into hierarchical RL algorithms (Barto and Mahadevan 2003, Sutton et al. 1999), but such models need only be limited, local models that focus on environmental regularities relevant to particular skills. On a deeper level, this view arises from the conviction that control rather than prediction must have played the dominant role in the evolution of cognition. The utility of prediction arises solely through its role in facilitating control. Although we cannot control the weather, we use weather predictions to control its impact on us: we cancel the picnic, carry an umbrella, etc. Of course, because prediction is so useful for control, we would expect intrinsic motivational mechanisms to exist that encourage accurate prediction, but according to the competence view this is not the sole underlying purpose of intrinsic motivation.[6]

5 Motivation and Reinforcement Learning

Although many computational models of RL contributed to how it is now studied in machine learning (e.g., Clark and Farley 1955, Mendel and Fu 1970, Mendel and McLaren 1970, Michie and Chambers 1968, Minsky 1954, Narendra and Thathachar 1989, Widrow et al. 1973), a most influential collection of ideas are those of A. H. Klopf—especially as it concerns the author and his students and our influence on RL. Klopf (1972, 1982) argued that homeostasis should not be considered the primary goal of behavior and learning, and that it is not a suitable organizing principle for developing artificial intelligence. Instead, he argued that organisms strive for a maximal condition:

> It will be proposed here that homeostasis is not the primary goal of more phylogenetically advanced living systems; rather, that it is a secondary or subgoal. It is suggested that the primary goal is a condition which ... will be termed heterostasis. An organism is said to be in a condition of heterostasis with respect to a specific internal variable when that variable has been maximized. (p. 10, Klopf 1982)

This proposal is built into all RL systems, where maximization—or more generally optimization—is the guiding principle instead of equilibrium-seeking. This basis of RL commonly generates several questions that deserve comment. First, what is the difference between maximization and equilibrium seeking? There is clearly no logical difference between maximizing and minimizing (one simply changes a sign: both are optimization), and is not an equilibrium-seeking

[6] Schmidhuber (2009) would argue that it is the other way around—that control is a result of behavior directed to improve predictive models. Resolution of this issue lies in the future.

system engaged in minimizing, specifically, minimizing the discrepancy between the current and the desired state? It is correct that equilibrium seeking involves minimizing, but it is a restricted variety of minimization based on assumptions that are not true in general. Consider the difference between searching for something that you will recognize when you find it, such as a specific web site, and searching for the "best" of something, such as (what you consider to be) the best-tasting wine. In the former case, when the desired item is found, the search stops, whereas in the latter, the search must go on—at least in principle—until you have sampled every variety of wine. The former case is like equilibrium seeking: the search is looking for zero discrepancy between the current and desired state, whereas the latter is like RL, where incessant exploration is called for.

A second question that comes up with regard to RL's focus on optimization is this one: Does this not conflict with the commonly made observation that nature does not seem to optimize, at either ontogenetic or phylogenetic scales? Since biological adaptation does not produce optimal solutions, any view of nature based on optimization must be incorrect. The answer to this is that adopting an optimization framework does not imply that the results are always optimal. Indeed, in many large-scale optimization problems, globally optimal results are hardly ever achieved, and even if they were, one would never know it. The focus is on the *process*, which involves incessant exploration, and not the desired outcomes—which are almost never achieved.

Its emphasis on optimization instead of equilibrium-seeking makes RL closer to hedonic views of motivation than to Hullian views. However, whereas a hedonic view of motivation is usually associated with affective mental states, RL—at least as described by Sutton and Barto (1998)—does not venture in to this territory. There is no mention of a "feeling" of pleasantness or unpleasantness associated with reward or penalty. This may be excluding an important dimension, but that dimension is not an essential component of a framework based on optimization. What is essential is RL's attitude toward the following questions put forward by Cofer and Appley (1964):

> Is organismic functioning conservative or growth-oriented? ... Does the organism's behavior serve primarily to maintain a steady state or equilibrium, that is, is it homeostatically organized and conservative? Or, alternatively, does the organism's behavior serve to take it to new levels of development, unfolding potentialities that a conservative principle could not encompass? (p. 15, Cofer and Appley 1964)

This non-conservative view of motivation, as contrasted with one based on maintaining equilibrium, is a good summary of what makes RL and other frameworks based on optimization attractive approaches to designing intelligent agents.

The importance of this for learning, especially machine learning, is that the learning algorithms most commonly studied are suitable only for supervised learning problems. As discussed above, these algorithms work by making adjustments directed toward eliminating mismatches between expectations

and what actually occurs. They are therefore essentially equilibrium-seeking mechanisms in the sense of attempting to zero out an error measure. Indeed, learning algorithms such as Rosenblatt's perceptron (Rosenblatt 1962) and Widrow and Hoff's ADALINE (Widrow and Hoff 1960), and their many descendants, explicitly employ the negative feedback principle of an equilibrium-seeking servo mechanism. An RL algorithm, in contrast, is attempting to extremize a quantity, specifically a measure of reward. Although supervised learning clearly has its place, the RL framework's emphasis on incessant activity over equilibrium-seeking makes it essential—in the author's opinion—for producing growth-oriented systems with open-ended capacities.[7]

6 Intrinsic Motivation in Reinforcement Learning

In the RL framework, an agent works to maximize a quantity based on an abstract reward signal that can be derived in many different ways. As emphasized in Section 1, the RL framework "does not care" where the reward signal comes from. The framework can therefore encompass both homeostatic theories of motivation in which rewards are defined as drive reduction, as has been done in many motivational systems for artificial agents (Savage 2000), and non-homeostatic theories that can account, for example, for the behavioral effects of electrical brain stimulation and addictive drugs. It can also include intrinsic reward signals.

In presenting the basic RL framework in Section 2 above, we emphasized that an RL agent should not be thought of as an entire animal or robot, and that the box labeled "Environment" in Figure 1 represents not only an animal's or robot's external world but also components within the animal or robot itself. Figure 2 is a refinement of Figure 1 that makes this explicit by dividing the environment into an *external environment* and an *internal environment*. The external environment represents what is outside of the animal or robot (which we will refer to as an "organism", as labeled in the figure), whereas the internal environment consists of components that are inside the organism.[8] Both components together comprise the environment of the RL agent.[9]

This refinement of the usual RL framework makes it clear that *all reward signals are generated within the organism*. Some reward signals may triggered by sensations produced by objects or events in the external environment, such as a pat on the head or a word of praise; others may be triggered by a combination of

[7] These comments apply to the "passive" form of supervised learning; not necessarily to the extension known as "active learning" (Settles 2009), in which the learning agent itself chooses training examples. Although beyond this chapter's scope, active supervised learning is indeed relevant to the subject of intrinsic motivation.

[8] We are relying on a common-sense notion of an organism's boundary with its external environment, recognizing that this may be not be easy to define.

[9] Figure 2 shows the organism containing a single RL agent, but an organism might contain many, each possibly having its own reward signal. Although not considered here, the multi-agent RL case (Busoniu et al. 2008) poses many challenges and opportunities.

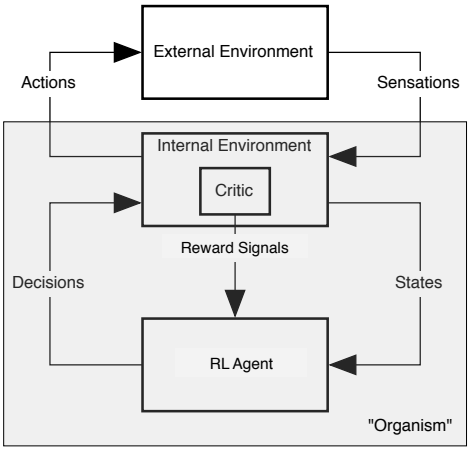

Figure 2. *Agent-Environment Interaction in RL. A refinement of Figure 1 in which the environment is divided into and internal and external environment, with all reward signals coming form the former. The shaded box corresponds to what we would think of as the "organism." Adapted from Barto et al. (2004).*

external stimulation and conditions of the internal environment, such as drinking water in a state of thirst. Still other reward signals may be triggered solely by activity of the internal environment, such as entertaining a thought or recalling a memory. All of these possibilities can be accommodated by the RL framework as long as one does not identify an RL agent with a complete organism.

It is tempting to directly connect the distinction between the external and internal environments with the distinction between extrinsic and intrinsic reward signals: extrinsic reward signals are triggered by objects or events in the external environment, whereas intrinsic reward signals are triggered solely by activity of the internal environment. Unfortunately, this view does not do justice to the complexity and variability of either extrinsically or intrinsically rewarding behavior.

According to Bindra's (1978) account mentioned in Section 3, for example, an organism's internal state, such as arousal level, blood-suger level, hormone levels, etc., interacts with features of an object or event signaled by external stimulation to generate a central motive state. Assuming this central motive state influences the generation of reward signals, this account clearly involves both the organism's internal and external environments. Even putting aside Bindra's account, it is clear that the state of an organism's internal environment modulates how external stimulation is transduced into reward signals. Moreover, for many instances of what we would consider intrinsically motivated activity, for example when behavior is the result of pure curiosity, an organism's

external environment is often a key player: objects and events in the external environment trigger curiosity, surprise, and other constituents of intrinsically motivated behavior.

Despite the difficulty of aligning extrinsic and intrinsic reward signals with and organism's external and internal environments, the internal environment may play a larger—or at least, a different—role in generating reward signals associated with intrinsic motivation. For example, a salient external stimulus might generate a reward signal to the degree that it is unexpected, where the expectancy is evaluated by processes in the internal environment and information stored there. Novelty, surprise, incongruity, and other features that have been hypothesized to underlie intrinsic motivation all depend on what the agent has already learned and experienced, that is, on its memories, beliefs, and internal knowledge state, all of which are components of the state of the organism's internal environment. One can think of these as the "informational", as opposed to the vegetative, aspects of the internal environment.

The approach to intrinsically motivated RL taken by the author and colleagues is to include these kinds of rewards as components of the RL agent's primary reward function. This is consistent with the large body of data alluded to above showing that intrinsically motivated behavior is not dependent on secondary reinforcement, that is, behavior is not intrinsically motivated because it had previously been paired with the satisfaction of a primary biological need in the animal's own experience (Deci and Ryan 1985). It is also in accord with Schmidhuber's (1991a, 2009) approach to curious RL systems where both normal and curious behavior use the same mechanism. *Some* behavior that we might call intrinsically motivated could be motivated through learned secondary reward signals, but this is not a necessary feature.

If the internal/external environment dichotomy does not provide a way to cleanly distinguish between extrinsic and intrinsic reward signals, what does? The author's current view is that there is no clean distinction between these types of reward signals; instead there is a continuum ranging from clearly extrinsic to clearly intrinsic. This view is the result of considering the issue from an evolutionary perspective, which is taken up next.

7 Evolutionary Perspective

Intrinsically motivated behavior is not anomalous from an evolutionary perspective. Intrinsically motivated behavior and what animals learn from it clearly contribute to survival and reproductive success. The success of humans owes a lot to our intrinsic urge to control our environments. It is not surprising, then, that machinery has evolved for ensuring that animals gain the kinds of experiences from which they can acquire knowledge and skills useful for survival and reproduction. Building in reward mechanisms to motivate knowledge-acquiring and skill-acquiring behavior is a parsimonious way of achieving this—enlisting motivational processes to appropriately direct behavior. From an evolutionary perspective, then, there is nothing particularly mysterious about intrinsically

motivated behavior. But can an evolutionary perspective help us understand the relationship between intrinsic and extrinsic motivation and reward signals?

Inspired in part by economists Samuelson and Swinkels (2006), who asked the question

> ... given that successful descendants are the currency of evolutionary success, why do people have utility for anything else? (p. 120, Samuelson and Swinkels 2006),

Singh et al. (2009, 2010) placed reward processes in an evolutionary context, formulating a notion of an *optimal reward function* given an evolutionary fitness function and a distribution of environments. Results of computational experiments suggest how both extrinsically and intrinsically motivated behaviors may emerge from such optimal reward functions. The approach taken in these studies was to evaluate entire primary reward functions in terms of how well simulated agents learning according to these reward functions performed as evaluated by a separate "evolutionary fitness function." An automated search in a space of primary reward functions could then be conducted to see which reward function would confer the most evolutionary advantage to the learning agent. Key to this approach is that each agent's behavior was evaluated across multiple environments, where some features remained constant across all the environments and others varied from environment to environment.

Readers should consult Singh et al. (2009, 2010) for details, but a main lesson from these studies is that the difference between intrinsic and extrinsic reward may be one of degree rather than one that can be rigorously defined by specific features. When coupled with learning, a primary reward function that rewards behavior that is *ubiquitously useful across many different environments* can produce greater evolutionary fitness than a function exclusively rewarding behavior directly related to the most basic requisites of survival and reproduction. For example, since eating is necessary for evolutionary success in all environments, we see primary reward signals generated by eating-related behavior. But reward functions that in addition reward behavior less directly related to basic needs, such as exploration and play, can confer greater evolutionary fitness to an agent. This is because what is learned during exploration and play contributes, within the lifetime of an animal, to that animal's ability to reproduce. It is therefore not surprising that evolution would give exploration, play, etc. positive hedonic valence, i.e., would make them rewarding.

A possible conclusion from this evolutionary perspective is that what we call extrinsically rewarding stimuli or events are those that have a relatively immediate and direct relationship to evolutionary success. What we call intrinsically rewarding activities, on the other hand, bear a much more distal relationship to evolutionary success. The causal chain from these behaviors to evolutionary success is longer, more complex, and less certain than the chain from what we typically call extrinsically motivated behavior. This makes it difficult to recognize evolutionarily beneficial consequences of intrinsically motivated behavior. Berlyne (1960) used the term "ludic behavior" (from the Latin *ludare*, to play) which "... can best be defined as any behavior that does not have a biological

function that we can clearly recognize." It is not clear that this property adequately characterizes all intrinsically motivated behavior, but it does capture something essential about it.

The relationship between intrinsic reward and evolutionary success is analogous to the relationship between learned, or secondary, reward and primary reward. In the latter case, a stimulus or activity comes to a generate reward signals to the extent that it predicts future primary reward. This is the basic mechanism built into RL algorithms that estimate value functions. Behavior is selected on the basis of predictions of the total amount of primary reward expected to accrue over the future, as represented by the learned value function (see Section 2 above). Through this process, good actions can be selected even when their influence on future primary reward is only very indirect. Imagine, for example, an early move in a game of backgammon that helps to set the stage for a much later advance, which ultimately results in winning the game. An RL algorithm such as the one used in the program TD-Gammon (Tesauro 1994) uses the value function it learns to effectively reward this move immediately when it is taken.

A similar situation occurs in the case of intrinsic reward, except that in this case a stimulus or activity comes to elicit a *primary* reward signal to the extent that it predicts eventual *evolutionary* success. In this case, the evolutionary process confers a rewarding quality to the stimulus or activity. Although this is clearly different from the what happens with secondary reward, where a stimulus becomes rewarding through learning that takes place within the lifetime of an individual animal, in both cases the rewarding quality arises due to a predictive relationship to a "higher level" measure of success: reproductive success in the case of evolution and primary reward in the case of secondary reinforcement.

This evolutionary context provides insight into the kinds of behavior we might expect an evolved reward function to encourage. We might expect a reward function to evolve that "taps into" features that were constant across many ancestral generations, but we would not expect one to evolve that exploits features that change from generation to generation. For example, if food tends to be found in places characterized by certain fixed features, we might expect a primary reward signal to be elicited by these features to encourage approach behavior. However, we would not expect approach to specific spatial locations to be rewarding unless these locations were the loci of sustenance for generation after generation. Learning can exploit features that maintain relatively fixed relationships to reward within a single agent's lifetime, whereas the evolutionary process is able to exploit larger-scale constancies that transcend individual agents and environments.

As a consequence, an animal's reward systems will promote behavior that is *ubiquitously useful across many different environments*. In some cases, this behavior's utility is easily recognizable and appears to be directed toward a proximal goal with obvious biological significance. In other cases, the behavior's utility is difficult to recognize because it contributes more indirectly and with less certainty to evolutionary success: its purpose or goal may be so far removed from the behavior itself that it may appear to have no clear purpose at all.

A somewhat similar relationship exists between basic and applied, or programatic, research. In arguing for the importance of basic research in his famous report to the United States president, Vannevar Bush (1945) wrote:

> Basic research is performed without thought of practical ends. It results in general knowledge and an understanding of nature and its laws. This general knowledge provides the means of answering a large number of important practical problems, though it may not give a complete specific answer to any one of them. The function of applied research is to provide such complete answers. The scientist doing basic research may not be at all interested in the practical applications of his work, yet the further progress of industrial development would eventually stagnate if basic scientific research were long neglected. ... Basic research leads to new knowledge. It provides scientific capital. It creates the fund from which the practical applications of knowledge must be drawn. (Bush 1945)

It is not misleading to think of basic research as intrinsically motivated, whereas applied research is extrinsically motivated, being directed toward a specific identifiable end. Bush was asserting that basic research has enormous practical utility, but it is not an immediate or certain consequence of the activity.

Although the distinction between basic and applied research seems clear enough, one may be hard pressed to point to features of specific research activities that would mark them, out of a broader context, as being conducted as part of a basic or an applied research project. The same seems true of intrinsically and extrinsically motivated behavior. The evolutionary perspective suggests that there are no hard and fast features distinguishing intrinsic and extrinsic reward. There is rather a continuum along which the directness of the relationship varies between sources of reward signals and evolutionary success. The claim here is that what we call intrinsically rewarding behavior is behavior that occupies the range of this continuum in which the relationship is relatively indirect. Whether it is direct of indirect, moreover, this relationship to evolutionary success is based on environmental characteristics that have remained relatively constant, though of varying reliability, over many generations.

This leads to a final observation that reconciles this view of intrinsic reward signals with others that have been put forward, e.g., by Oudeyer and Kaplan (2007). Prominent among environmental features that maintain a relatively constant relationship to evolutionary success are features of the internal portion of an organism's environment, as depicted in Figure 2. What is labelled there the internal environment is carried along in relatively unchanging form from generation to generation. Therefore we would expect an animal's primary reward function to encourage a variety of behaviors that involve features of this part of the learning system's environment. This would include behaviors that we think of as involving curiosity, novelty, surprise, and other internally-mediated features usually associated with intrinsic reward. Thus, in addition to suggesting why it seems so difficult to place the distinction between intrinsic and extrinsic reward on a rigorous footing, the evolutionary perspective suggests an

explanation for why the prototypical examples of activities that we think of as intrinsically rewarding tend to heavily depend on variables that describe aspects of an animal's internal environment.

There is clearly much more to understand about the relationship between evolution and learning, and there is a large literature on the subject. Less has been written about the evolution of reward structures, though a number of computational studies have been published (Ackley and Littman 1991b, Damoulas et al. 2005, Elfwing et al. 2008, Littman and Ackley 1991, Schembri et al. 2007, Snel and Hayes 2008, Uchibe and Doya 2008). In addition to Singh et al. (2009, 2010), the most relevant computational study of which the author is aware is that of Ackley and Littman (1991a). Sorg, Singh, and Lewis (2010) provide computational results to support the view that a key role played by reward functions is to attenuate the negative consequences of various types of agent limitations, such as lack of information, lack of adequate time to learn, or lack of efficient learning mechanisms. This view is critical to reaching a better understanding of intrinsic motivation, and it is consistent with observations from economists who study the *evolution of preferences* in the context of game theory (Samuelson 2001).

8 Summary and Prospects

This chapter focuses on review and perspective, while saying little about architectures and algorithms for intrinsically motivated artificial agents. However, some general conclusions are supported by the views presented here that can help guide the development of competently autonomous artificial agents.

1. *RL is particularly suited for incorporating principles of motivation into artificial agents, including intrinsic motivation.* This chapter argues that an approach to building intelligent agents based on principles of optimization, instead of solely equilibrium-seeking, gives agents the kind of incessant activity that is essential for growth-oriented systems with open-ended capacities. A base in optimization does not mean that optimal solutions need ever be found: it is the process that is important.

2. *The distinction between an RL agent and its environment at the base of the RL formulation has to be looked at in the right way.* An RL agent is not the same as an entire animal or robot. Motivational processes involve state components that are internal to the animal or robot while at the same time being external to the RL agent. The sources of reward signals are, as is usual in the RL framework, external to the RL agent (so it cannot exert complete control over them), while still being within the animal or robot.

3. *State components that influence an RL agent's reward signals can include features of a robot's memories and beliefs in addition to "vegetative" features.* This follows from item (2) since this information is part of the RL

agent's environment. In fact, a robot's current policy, value function, and environment model are all possible influences on reward signals since they can also be components of the state of the RL agent's environment. This opens the possibility for defining many interesting reward functions.

4. *The view that motivation can be equated with the nature of an RL reward function is only part of the story.* In a reward-maximizing framework, there is a natural correspondence between the reward function and the forces that direct agent activity. However, this does not imply that the nature of a reward function accounts for all aspects of an agent's motivations. In both modeling biological agents and building artificial agents, other components are important as well. For example, there are prominent roles for complex structures of built-in behaviors, and there may be multiple optimizing components with different objectives, requiring arbitration mechanisms to coordinate among competing goals. Further, theories relevant to intrinsic motivation have been proposed that are not based on RL, e.g., Andry et al. (2004), Baranes and Oudeyer (2010), Friston et al. (2010), Hesse et al. (2009).

5. *There is no hard-and-fast distinction between extrinsic and intrinsic reward signals.* There is rather a continuum along which reward signals fall, ranging from signals clearly related to proximal goals with obvious biological utility to signals with less direct and less reliable biological utility. These latter signals underlie what we think of as intrinsically motivated behavior. This view is suggested by recent computational study by Singh, Lewis, and Barto (2009, 2010), which explores the concept of evolutionarily optimal reward functions as discussed in Section 7.

6. *Despite the difficulty in giving the extrinsic/extrinsic distinction a completely satisfactory formal definition, the distinction is still useful.* In particular, the psychologist's definition, where extrinsic motivation means doing something because of some specific rewarding outcome, and intrinsic motivation means "doing something because it is inherently interesting or enjoyable" (Ryan and Deci 2000), is adequate for most purposes. It alerts us to the possible benefits of defining reward functions that depend on a wider range of factors than those usually considered in RL. Specifically, reward functions can depend on the state of a robot's internal environment, which includes remembered and learned information.

7. *It is not likely that there is a single unitary principle underlying intrinsic motivation.* Although the evolutionary perspective presented here does not give detailed information about what architectures and algorithms we should develop to produce intrinsically motivated artificial agents, it does suggest that the best reward functions will depend on the distribution of tasks at which we wish the agent to excel. Therefore, although some principles are undoubtedly widely-applicable—such as some of those already receiving attention in the literature—skepticism is justified about

the proposition that one principle suffices to account for all aspects of intrinsic motivation.

8. *Analogs of intrinsic motivation are destined to play important roles in future machine learning systems.* In the same way that intrinsic motivation plays a crucial role in directing human behavior for both children and adults, we can expect computational analogs to be important for directing the behavior of machine learning systems that are able to exert some control over their input by being embedded in physical or virtual worlds. Moreover, the progress occurring in making computational power a ubiquitous resource means that learning systems can be constantly active, even when they are not engaged in solving particular problems. Intrinsic motivation is the means for making the most of such idle times in preparation for problems to come. In the introduction to his 1960 treatise on curiosity, Berlyne wrote the following:

   > Until recently, rather little has been done to find out how animals behave, whether in the wild or in captivity, when they have nothing particular to do. (p. 1, Berlyne 1960)

   Although we may know—or at least hope we know!—what our computers are doing when they have nothing particular to do, it is clear that, like animals, they could be working to build the competencies needed for when they are called to action.

9 Conclusion

Building intrinsic motivation into artificial agents may bring to mind all the warnings from science fiction about the dangers of truly autonomous robots. But there are good reasons for wanting artificial agents to have the kind of broad competence that intrinsically motivated learning can enable. Autonomy is increasingly becoming a more common property of automated systems since it allows them to successfully operate for extended periods of time in dynamic, complex, dangerous environments about which little *a priori* knowledge is available. As automated systems inevitably assume more "unpluggable" roles in our lives, *competent autonomy* is becoming increasingly essential to prevent the kind of catastrophic break-downs that threaten our society. In a real sense, we already depend on systems, such as the power grid, that are essentially autonomous but seriously lacking in competence. Providing them with intrinsic motivations carefully crafted to embody desirable standards may be a path toward making artificial agents competent enough to fulfill their promise of improving human lives.

Acknowledgements

The author thanks Satinder Singh and Rich Lewis for developing the evolutionary perspective on this subject, Jonathan Sorg, for his important insights, and

colleagues Sridhar Mahadevan and Rod Grupen, along with current and former members of the Autonomous Learning Laboratory who have participated in discussing intrinsically motivated reinforcement learning: Will Dabney, Jody Fanto, Scott Kuindersma, George Konidaris, Scott Niekum, Özgür Şimşek, Andrew Stout, Chris Vigorito, and Pippin Wolfe. The author also thanks Pierre-Yves Oudeyer for his many helpful suggestions, especially regarding non-RL approaches to intrinsic motivation. Some of the work described here was supported by the National Science Foundation under Grant No. IIS-0733581 and by the Air Force Office of Scientific Research under grant FA9550-08-1-0418. Any opinions, findings, conclusions or recommendations expressed here are those of the author and do not necessarily reflect the views of the sponsors.

References

D. H. Ackley and M. Littman. Interactions between learning and evolution. *Artificial Life II, SFI Studies in the Sciences of Complexity*, X, 1991a.

D. H. Ackley and M. Littman. Interactions between learning and evolution. In C.G. Langton, C. Taylor, C.D. Farmer, and S. Rasmussen, editors, *Artificial Life II (Proceedings Volume X in the Santa Fe Institute Studies in the Sciences of Complexity*, pages 487–509. Addison-Wesley, Reading, MA, 1991b.

P. Andry, P. Gaussier, J. Nadel, and B. Hirsbrunner. Learning invariant sensorimotor behaviors: A developmental approach to imitation mechanisms. *Adaptive Behavior*, 12:117–140, 2004.

H. R. Arkes and J. P. Garske. *Psychological Theories of Motivation.* Brooks/Cole, Monterey CA, 1982.

A. Baranes and P.-Y. Oudeyer. Intrinsically motivated goal exploration for active motor learning in robots: A case study. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*, 2010.

A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamcal Systems: Theory and Applications*, 13: 341–379, 2003.

A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:835–846, 1983. Reprinted in J. A. Anderson and E. Rosenfeld (eds.), *Neurocomputing: Foundations of Research*, pp. 535-549, MIT Press, Cambridge, MA, 1988.

A. G. Barto, S. Singh, and N. Chentanez. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the International Conference on Developmental Learning (ICDL)*, 2004.

R. C. Beck. *Motivation. Theories and Principles, 2nd Edition.* Prentice-Hall, Englewood Cliffs NJ, 1983.

D. E. Berlyne. A theory of human curiosity. *British Journal of Psychology*, 45: 180–191, 1954.

D. E. Berlyne. *Conflict, Arousal. and Curiosity.* McGraw-Hill, N.Y., 1960.

D. E. Berlyne. *Aesthetics and Psychobiology.* Appleton-Century-Crofts, N.Y., 1971.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming.* Athena Scientific, Belmont, MA, 1996.

D. Bindra. How adaptive behavior is produced: A perceptual-motivational alternative to response reinforcement. *Behavioral and Brain Sciences*, 1:41–91, 1978.

C. Breazeal, A. Brooks, J. Gray, G. Hoffman, J. Lieberman, H. Lee, A. Lockerd, and D. Mulanda. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robotics*, 1, 2004.

V. Bush. Science the endless frontier: Areport to the president. Technical report, 1945.

L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews*, 38(2):156172, 2008.

W. B. Cannon. *The Wisdom of the Body.* W. W. Norton, New York, 1932.

W. A. Clark and B. G. Farley. Generalization of pattern recognition in a self-organizing system. In *Proceedings of the 1955 Western Joint Computer Conference*, pages 86–91, 1955.

C. N. Cofer and M. H. Appley. *Motivation: Theory and Research.* Wiley, New York, 1964.

T. Damoulas, I. Cos-Aguilera, G. M. Hayes, and T. Taylor. Valency for adaptive homeostatic agents: Relating evolution and learning. In M. S. Capcarrere, A. A. Freitas, P. J. Bentley, C. G. Johnson, and J. Timmis, editors, *Advances in Artificial Life: 8th European Conference, ECAL 2005, LNAI vol. 3636*, pages 936–945. Springer-Verlag, Berlin, 2005.

N. D. Daw and D. Shohamy. The cognitive neuroscience of motivation and learning. *Social Cognition*, 26(5):593–620, 2008.

P. Dayan. Motivated reinforcement learning. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Conference*, pages 11–18, Cambridge MA, 2001. MIT Press.

E. L. Deci and R. M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. Plenum Press, N.Y., 1985.

W. N. Dember and R. W. Earl. Analysis of exploratory, manipulatory, and curiosity behaviors. *Psychological Review*, 64:91–96, 1957.

W. N. Dember, R. W. Earl, and N. Paradise. Response by rats to differential stimulus complexity. *Journal of Comparative and Physiological Psychology*, 50:514–518, 1957.

A. Dickinson and B. Balleine. The role of leaning in the operation of motivational systems. In R. Gallistel, editor, *Handbook of Experimental Psychology 3rd Edition: Learning, Motivation, and Emotion*, pages 497–533. Wiley, New York, 2002.

S. Elfwing, E. Uchibe, K. Doya, and H. I. Christensen. Co-evolution of shaping rewards and meta-parameters in reinforcement learning. *Adaptive Behavior*, 16:400–412, 2008.

A. Epstein. Instinct and motivation as explanations of complex behavior. In D. W. Pfaff, editor, *The Physiological Mechanisms of Motivation*. Springer, New York, 1982.

K. J. Friston, J. Daunizeau, J. Kilner, and S. J. Kiebel. Action and behavior: A free-energy formulation. *Biological Cybernetics*, 2010. Pubished on line Feb. 11, 2020.

K. Groos. *The Play of Man*. D. Appleton, N.Y., 1901.

H. F. Harlow. Learning and satiation of response in intrinsically motivated complex puzzle performance by monkeys. *Journal of Comparative and Physiological Psychology*, 43:289–294, 1950.

H. F. Harlow, M. K. Harlow, and D. R. Meyer. Learning motivated by a manipulation drive. *Journal of Experimental Psychology*, 40:228–234, 1950.

D. O. Hebb. *The Organization of Behavior*. Wiley, N.Y., 1949.

I. Hendrick. Instinct and ego during infancy. *Psychoanal. Quart.*, 11:33–58, 1942.

F. Hesse, R. Der, M. Herrmann, and J. Michael. Modulated exploratory dynamics can shape self-organized behavior. *Advances in Complex Systems*, 12 (2):273–292, 2009.

C. L. Hull. *Principles of Behavior*. D. Appleton-Century, NY, 1943.

C. L. Hull. *Essentials of Behavior*. Yale University Press, New Haven, 1951.

C. L. Hull. *A Behavior System: An Introduction to Behavior Theory Concerning the Individual Organism*. Yale University Press, New Haven, 1952.

G. A. Kimble. *Hilgard and Marquis' Contitioning and Learning*. Appleton-Century-Crofts, Inc., New York, 1961.

S. B. Klein. *Motivation. Biosocial Approaches*. McGraw-Hill, New York, 1982.

A. H. Klopf. Brain function and adaptive systems—A heterostatic theory. Technical Report AFCRL-72-0164, Air Force Cambridge Research Laboratories, Bedford, MA, 1972. A summary appears in *Proceedings of the International Conference on Systems, Man, and Cybernetics*, 1974, IEEE Systems, Man, and Cybernetics Society, Dallas, TX.

A. H. Klopf. *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*. Hemisphere, Washington, D.C., 1982.

D. B. Lenat. *AM: An Artificial Intelligence Approach to Discovery in Mathematics*. PhD thesis, Stanford University, 1976.

D. J. Linden. *The Compass of Pleasure: How Our Brains Make Fatty Foods, Orgasm, Exercise, Marijuana, Generosity, Vodka, Learning, and Gambling Feel So Good*. Viking, New York, 2011.

M. L. Littman and D. H. Ackley. Adaptation in constant utility nonstationary environments. In *Proceedings of the Fourth International Conference on Genetic Algorithms*, pages 136–142. 1991.

M. Lungarella, G. Metta, R. Pfeiffer, and G. Sandini. Developmental robotics: A survey. *Connection Science*, 15:151–190, 2003.

N. J. Mackintosh. *Conditioning and Associative Learning*. Oxford University Press, New York, 1983.

D. McFarland and T. Bösser. *Intelligent Behavior in Animals and Robots*. MIT Press, Cambridge MA, 1993.

J. M. Mendel and K. S. Fu, editors. *Adaptive, Learning, and Pattern Recognition Systems: Theory and Applications*. Academic Press, New York, 1970.

J. M. Mendel and R. W. McLaren. Reinforcement learning control and pattern recognition systems. In J. M. Mendel and K. S. Fu, editors, *Adaptive, Learning and Pattern Recognition Systems: Theory and Applications*, pages 287–318. Academic Press, New York, 1970.

D. Michie and R. A. Chambers. BOXES: An experiment in adaptive control. In E. Dale and D. Michie, editors, *Machine Intelligence 2*, pages 137–152. Oliver and Boyd, Edinburgh, 1968.

M. L. Minsky. *Theory of Neural-Analog Reinforcement Systems and its Application to the Brain-Model Problem*. PhD thesis, Princeton University, 1954.

M. L. Minsky. Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers*, 49:8–30, 1961. Reprinted in E. A. Feigenbaum and J. Feldman, editors, *Computers and Thought*. McGraw-Hill, New York, 406–450, 1963.

S. O. Mollenauer. Shifts in deprivations level: Different effects depending on the amount of preshift training. *Learning and Motivation*, 2:58–66, 1971.

K. Narendra and M. A. L. Thathachar. *Learning Automata: An Introduction*. Prentice Hall, Englewood Cliffs, NJ, 1989.

J. Olds and P. Milner. Positive reinforcement produced by electrical stimulation of septal areas and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, 47:419–427, 1954.

P.-Y. Oudeyer and F. Kaplan. What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, 2007.

P.-Y. Oudeyer, F. Kaplan, and V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11, 2007.

H. L. Petri. *Motivation: Theory and Research*. Wadsworth Publishing Company, Belmont CA, 1981.

J. Piaget. *The Origins of Intelligence in Children*. Norton, N.Y., 1952.

R. W. Picard. *Affective Computing*. MIT Press, Cambridge MA, 1997.

C. G. Prince, Y. Demiris, Y. Marom, H. Kozima, and C. Balkenius, editors. *Proceedings of the Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems. Lund University Cognitive Studies Volume 94*. Lund University, Lund, Sweden, 2001.

R. A. Rescorla and A. R. Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black and W. F. Prokasy, editors, *Classical Conditioning II*, pages 64–99. Appleton-Century-Crofts, New York, 1972.

F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, DC, 1962.

R. M. Ryan and E. L. Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25:54–67, 2000.

L. Samuelson. Introduction to the evolution of preferences. *Journal of Economic Theory*, 97:225–230, 2001.

L. Samuelson and J. Swinkels. Information, evolution, and utility. *Theoretical Economics*, 1:119–142, 2006.

T. Savage. Artificial motives: A review of motivation in artificial creatures. *Connection Science*, 12:211–277, 2000.

M. Schembri, M. Mirolli, and G. Baldassarre. Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In *Proceedings of the 6th International Conference on Development and Learning (ICDL2007)*, 2007.

J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pages 222–227, Cambridge, MA, 1991a. MIT Press.

J. Schmidhuber. Adaptive confidence and adaptive curiosity. Technical Report FKI-149-91, Institut für Informatik, Technische Universität München, Arcisstr. 21, 800 München 2, Germany, 1991b.

J. Schmidhuber. What's interesting? Technical Report TR-35-97, IDSIA, Lugano, Switzerland, 1997.

J. Schmidhuber. Artificial curiosity based on discovering novel algorithmic predictability through coevolution. In *Proceedings of the Congress on Evolutionary Computation*, volume 3, pages 1612–1618. IEEE Press, 1999.

J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In G. Pezzulo, M. V. Butz, O. Sigaud, and G. Baldassarre, editors, *Anticipatory Behavior in Adaptive Learning Systems. From Psychological Theories to Artificial Cognitive Systems*, pages 48–76. Springer, Berlin, 2009.

W. Schultz. Reward. *Scholarpedia*, 2(3):1652, 2007a.

W. Schultz. Reward signals. *Scholarpedia*, 2(6):2184, 2007b.

B. Settles. Active learning literature survey. Technical Report 1648, Computer Sciences, University of Wisconsin–Madison, Madison, WI, 2009.

S. Singh, A. G. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, Cambridge MA, 2005. MIT Press.

S. Singh, R. L. Lewis, and A. G. Barto. Where do rewards come from? In N.A. Taatgen and H. van Rijn, editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2601–2606. Cognitive Science Society, 2009.

S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):7082, 2010. Special issue on Active

Learning and Intrinsically Motivated Exploration in Robots: Advances and Challenges.

M. Snel and G. M. Hayes. Evolution of valence systems in an unstable environment. In *Proceedings of the 10th International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, pages 12–21, Osaka, Japan, 2008.

J. Sorg, S. Singh, and R. L. Lewis. Internal rewards mitigate agent boundedness. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning*, pages 1007–1014, 2010.

R. S. Sutton. Reinforcement learning architectures for animats. In *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pages 288–296, Cambridge, MA, 1991. MIT Press.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

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

G. J. Tesauro. TD–gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215–219, 1994.

A. L. Thomaz and C. Breazeal. Transparency and socially guided machine learning. In *Proceedings of the 5th International Conference on Developmental Learning (ICDL)*, 2006.

A. L. Thomaz, G. Hoffman, and C. Breazeal. Experiments in socially guided machine learning: Understanding how humans teach. In *Proceedings of the 1st Annual conference on Human-Robot Interaction (HRI)*, 2006.

E. L. Thorndike. *Animal Intelligence*. Hafner, Darien, Conn., 1911.

F. M. Toates. *Motivational Systems*. Cambridge University Press, Cambridge, 1986.

E. C. Tolman. *Purposive Behavior in Animals and Men*. Naiburg, N.Y., 1932.

R. Trappl, P. Petta, and S. Payr, editors. *Emotions in Humans and Artifacts*. MIT Press, Cambridge MA, 1997.

E. Uchibe and K. Doya. Finding intrinsic rewards by embodied evolution and constrained reinforcement learning. *Neural Networks*, 21(10):1447–1455, 2008.

M. D. Waltz and K. S. Fu. A heuristic approach to reinforcment learning control systems. *IEEE Transactions on Automatic Control*, 10:390–398, 1965.

J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen. Autonomous mental development by robots and animals. *Science*, 291:599–600, 2001.

P. J. Werbos. Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research. *IEEE Transactions on Systems, Man, and Cybernetics*, 17:7–20, 1987.

R. W. White. Motivation reconsidered: The concept of competence. *Psychological Review*, 66:297–333, 1959.

B. Widrow and M. E. Hoff. Adaptive switching circuits. In *1960 WESCON Convention Record Part IV*, pages 96–104, NY, 1960. Institute of Radio Engineers. Reprinted in J. A. Anderson and E. Rosenfeld, *Neurocomputing: Foundations of Research*, pp. 126–134. MIT Press, Cambridge, MA, 1988.

B. Widrow, N. K. Gupta, and S. Maitra. Punish/reward: Learning with a critic in adaptive threshold systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 3:455–465, 1973.

P. T. Young. Hedonic organization and regulation of behavior. *Psychological Review*, 73:59–86, 1966.