Review Article **Open Access**

Nazmul Siddique*, Paresh Dhakan, Inaki Rano, and Kathryn Merrick

A Review of the Relationship between Novelty, Intrinsic Motivation and Reinforcement Learning

https://doi.org/10.1515/pjbr-2017-0004
Received May 3, 2016; accepted October 4, 2017

Abstract: This paper presents a review on the tri-partite relationship between novelty, intrinsic motivation and reinforcement learning. The paper first presents a literature survey on novelty and the different computational models of novelty detection, with a specific focus on the features of stimuli that trigger a Hedonic value for generating a novelty signal. It then presents an overview of intrinsic motivation and investigations into different models with the aim of exploring deeper co-relationships between specific features of a novelty signal and its effect on intrinsic motivation in producing a reward function. Finally, it presents survey results on reinforcement learning, different models and their functional relationship with intrinsic motivation.

Keywords: Novelty, intrinsic motivation, reinforcement learning, reward function, habituation, action learning

1 Introduction

There are many controversies on an acceptable definition of novelty. There have been debates on whether it is a perception about object, event, act, or concept and how such phenomenon can be described. These are the fundamental questions posed by many researchers. Philosophers, psychologists and scientists have been trying to answer such questions for a long time. However, as yet there is no strict definition of novelty. As the tools and techniques used by these groups of researchers are different, they differ in their definition, argument and proofs. While one group argues novelty as the difference between perceptions at two different time instants, the others regarded it as the process of identifying stimuli that are different [1–3]. The novelty which the scientists are trying to characterise is mainly physical (i.e. visual) or behavioural novelty. We are not talking about the chemistry of novelty, e.g. taste or smell, as the tools for such novelty are limited. In all cases, the perception of novelty is in the brain of human and animal whether it is a physical, visual or behavioural novelty. On the occurrence of a physical or behavioural event, it stimulates the brain activity. The brain accepts the event as novel and slowly gets used to it. This phenomenon is known as habituation. Over the course of habituation, average responses to the event decrease and the event is no longer novel. There have been numerous methods on novelty detection (ND) reported in the literature [1, 2, 4–11].

Motivation has been a research topic for philosophers and psychologist for a long time. There is reason behind every activity being performed by humans and animals. Reason is to minimise a need whether it is biological or psychological (i.e. mental). An interesting question posed by many researchers is why infants and children engage in activities. They engage in activities of various types out of curiosity or fun without being rewarded. This kind of behaviour is also observable in adults and animals. Such behavioural phenomena are termed as intrinsic motivation (IM). Psychologists differentiate between intrinsic and extrinsic motivation. When the reward comes from the environment, it is termed as extrinsic motivation [12]. The inherent mechanism of IM in humans and animals is still unknown. The psychologists have been trying to understand IM and to explain the behaviourist theory of the learning and drives [12–14]. Some researchers defined IM as doing an activity for its satisfaction, for improving knowledge or for improving competency or skill. Psychologists are trying to develop a behaviourist theory while scientists in the cognitive computing have been trying to develop models of IM in quest of making intelligent machines with human-like behaviour. There have been different approaches in modelling IM. Again, researchers are divided on the issue of instigating signal for the motivation to happen. Some re-

***Corresponding Author: Nazmul Siddique:** Intelligent Systems Research Centre, Ulster University, UK, Email: nh.siddique@ulster.ac.uk
Paresh Dhakan, Inaki Rano: Intelligent Systems Research Centre, Ulster University, UK, Email: dhakan-p@ulster.ac.uk, i.rano@ulster.ac.uk
Kathryn Merrick: Department of Engineering and Information Technology, University of New South Wales, Australia, E-mail: k.merrick@adfa.edu.au

searchers argue that the signal comes from the internal mechanism of the brain while others are saying that the signal is initiated from the outside world. The most influential proposal was by Berlyne [13] who suggested that the IM effects involve novelty.

Reinforcement learning (RL) is the process of improving performance through trial-and-error experience. The simplest RL is based on the common-sense idea that if an action is followed by an improvement or satisfaction in the state of affairs, then the tendency to produce that action is strengthened (i.e. reinforced). This is known as Thorndike's [15] "law of effect." The reinforcement comes as a reward or punishment. Berlyne [16] subsequently revealed and highlighted how properties of stimuli can drive exploration and guide learning when stimuli are complex, unexpected or surprising. In a study of artificial creativity, Saundars and Gero [17] developed a system that demonstrates intrinsic motivation based on novelty and surprise following Berlyne's theories.

RL algorithm has been in existence since the 1990s. IM comes from developmental psychology into cognitive computing where IM is used as an internal reward mechanism that fits well with a simple generic RL algorithm. Developmental psychologists also observed that it is novelty that attracts infants to engage in activities and learning. There have been numerous reports in the literature on the three entities: ND, IM and RL. From these studies, a three-way relationship can be recapitulated between ND, IM and RL. This relationship can help in development of an integrated model for an autonomous learning system. This paper, therefore, will first present a review of the current literature and the available models for ND, IM and RL. Secondly, it will present a discussion on the functional correlation between the three constituent parts leading to an integrated model.

The paper is organised as follows. Section 2 will present results of the literature survey on novelty and the different computational models of novelty detection with a specific focus on the features of stimuli that trigger a Hedonic value for generating novelty signal and features of novelty signals. Section 3 will present the results of the investigation into intrinsic motivation and the different computational models with the aim of seeking deeper co-relationship between specific features of novelty signal and its effect on intrinsic motivation in producing a reward function. Section 4 will present the results of a survey into reinforcement learning and its relationship with intrinsic motivation. The aim of this section is to investigate the functional relationship between intrinsic motivation and reinforcement learning. Therefore, the focus of this section will be to analyse the specific features of reward function and its influence on the learning activity and continuity of the course of action. The final section 5 will recapitulate the findings from the preceding sections to establish the acclaimed functional co-relation between the three constituent parts and presentation of an integrated model. Some concluding remarks on the investigation and the relationship will be made in section 6.

2 Novelty

In general, novelty may be attributed to the features of an object or an event with the object in the forefront or an act of manipulating an object. Although there is no strict definition of novelty, it is widely regarded [1–3] as the process of identifying novel stimuli that are different from anything known before, new, interesting and often seeming slightly strange [18]. There have been numerous methods on novelty detection reported in the literature [1, 2, 4–7, 10, 11, 19]. Based on the definitions above, a number of novelty detection methods have been proposed in the literature, mainly focusing on novelty detection, outlier detection and anomaly detection. The three broad categories are discussed in brief.

2.1 Novelty Detection

Two main categories of novelty detection methods: statistical and neural network based approaches are distinguished by researchers [1, 6, 7]. Statistical based methods typically test samples whether they come from the same distribution or not based on prior assumptions on the data distributions, e.g. Gaussian, while the neural network based approaches map the data to different classes, e.g. known and unknown classes [7]. Regularization, pruning and constructive algorithms are common in neural network-based approaches. Constructive algorithms are usually preferred as they are advantageous to start with a smaller network that trains faster and can be expanded as the training progresses.

A similar approach to novelty detection is outlier detection defined by Grubbs [20] that states "an outlier is one that appears to deviate distinctly from other members of the sample". Three types of approaches are distinguished. The first types of outliers are determined without any prior knowledge of the data. The second types of outliers involve normal and abnormal data to be determined using supervised classification. The third types mainly model normal cases together with a few abnormal cases. Hodge and Aus-

tin [6] acknowledged the third type as novelty detection or novelty recognition. Anomaly detection determines patterns in data that do not conform to a well defined normal behaviour [4]. Anomalies can be classified into three categories. The first category includes point anomalies, which occur when an individual data instance is considered anomalous with respect to the rest of the data. The second category is contextual anomalies, which include data instances that are considered anomalous in a particular context. The third category is collective anomalies, which consist of sets of data instances that are anomalous with respect to the entire data set. Data boundaries between normal data and anomalies or between known and unknown data (i.e. novelties in terms of novelty detection) are the issues associated with anomaly detection. An acute issue is the noise in the data that seem similar to actual anomalies and makes identification of novelty difficult.

In all of these methods discussed above, data are presented as stimuli to a novelty detection system. Based on the data distribution, also to be termed as class or pattern, the novelty detection system produces a novelty signal. A conceptual diagram is present in Figure 1(a). A generic term stimulus is used to represent the data distribution, class or pattern. Stimuli, data distribution, class or pattern will be used interchangeably throughout the paper.

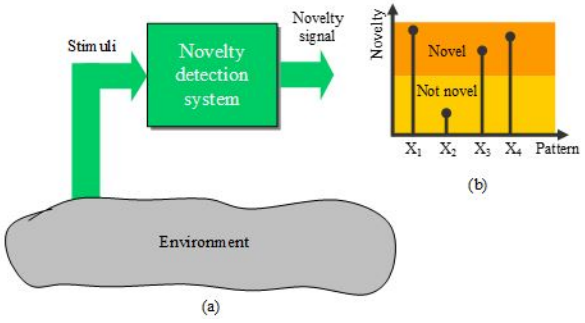

Figure 1: Novelty systems: (a) novelty signal as a function of stimuli; (b) discrete novel signal.

The important thing here is to observe the novelty signal and its characteristic features. In the first instance, it appears that the novelty signal is discrete and binary. That is, on presentation of stimuli or pattern, novelty detection (ND) system identifies the pattern as novel or not novel when data distribution (class or pattern) goes beyond a threshold value as illustrated in Figure 1(b). Also, it has been observed in human and animals and also verified experimentally by psychologists that repeated exposure to stimuli decreases responsiveness [21, 22] and over a period of time the repeating stimuli is no longer novel. That is, the novelty signal is continuous, gradually increases with time and reaches its highest value that continues for a certain period depending on the complexity of the stimuli and then decreases. This behavioural phenomenon is called habituation [23, 24]. Sokolov [25] pointed out that habituation relates to a function that declines in response to a feature when that feature is no longer novel and has no rewarding consequences. This view of habituation is congruent with Schmidhuber's [19] compression progress notion as it does not allow for further compression progress when data are no longer novel.

2.2 Novelty and habituation

In general, habituation is defined as a reversible decrease of response to repeated stimulation. It is a widely studied type of behavioural plasticity [26]. The characteristic features of habituation as identified by researchers [21, 23, 24, 26] are: exponential decrease of the response with repeated exposure to a stimulus, faster habituation with lower intensity stimuli, faster habituation with shorter intervals between stimuli, spontaneous recovery of the response after a period of break and recovery of the response or dishabituation of the stimuli previously habituated upon the introduction of a new stimulus. Various models have been proposed that reproduce the characteristic features of habituation [26–30]. In simple words, any stimulus, be it visual, auditory or touch sensed through sensors, will produce a data distribution (class or pattern). Data from auditory or mechanoreceptors may require further transformation and/or scaling to be detected as novel. If it is novel, it will habituate over time.

The well-known model of habituation and dishabituation is the one suggested by Stanley [26] who used the first-order differential equation defined by equation Eq. (1).

$$\tau_i \frac{h_i(t)}{dt} = \alpha \left[h_0 - h_i(t)\right] - S(t), \tag{1}$$

where h_0 is the initial value of the habituation level, $h_i(t)$ is the current habituation level, τ_i and α are time constants that control the habituation rate and the recovery rate respectively. $S(t)$ is the external stimulus defined by $S(t) = e^{\|X - w_s\|}$, where X is the input pattern and w_s is the weight vector. As the habituation value has a bounded output, it can be neatly used as a measure of the degree of novelty for any input. A generic behaviour of Stanley's model (habituation curve) is graphically shown in Figure 2.

The values of habituation are given by the equation range between [0, 1]. The value of 1 signifies a low value for

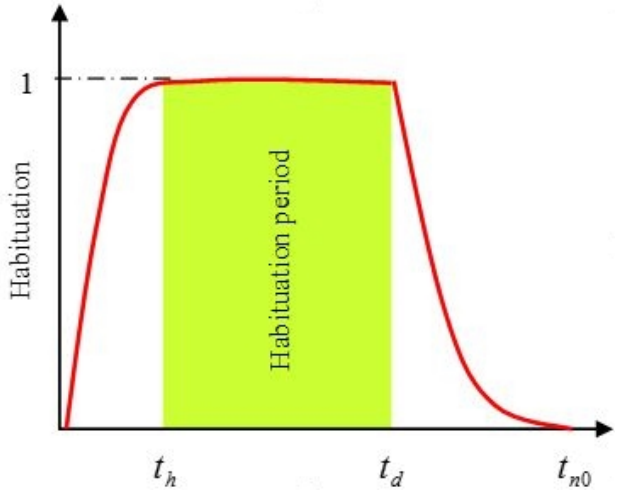

Figure 2: Habituation curve representing novelty over time.

habituation. That means the stimuli or perception has the maximum novelty. Some examples of habituation function characterising the features discussed above are shown in Figure 3.

During repeated testing, human (specifically infants) habituate to test items [24]. Figure 3(a) shows the habituation for repeated regular patterns and exponential decrease after habituation is complete. This might vary as a function of the degree of novelty. The duration $(t_d - t_h)$ will be minimum, though it may vary from person to person. It was found in an experiment by Hunter et al. [31] that infants look longer at stimuli that are relatively complex. Figures 3(b)-(c) show the habituation process of complex patterns. The situation can be described as habituation of random patterns, which takes longer but may fail to succeed due to the randomness of patterns. In these cases, the duration $(t_d - t_h)$ will be prolonged and may vary from person to person.

Novelty and habituation are closely related to attention. Attention is attracted by novel stimuli and similar to habituation, attention is not constant over time. Valasco et al. [32] investigated various somatic evoked responses which showed that there is a direct correlation between the somatic evoked response amplitude and novelty, habituation, attention and distraction. They found that somatic evoked response amplitude decreased when subjects shifted from novelty to habituation and increased from habituation to attention. Hutt [33] also reported the correlation between novelty and attention by empirical investigations on children. A recent review on the state of the art of attention mechanisms can be found in Ferreira et al. [34].

Stanley's computational model of habituation [26] has also been used in novelty detection by many research-

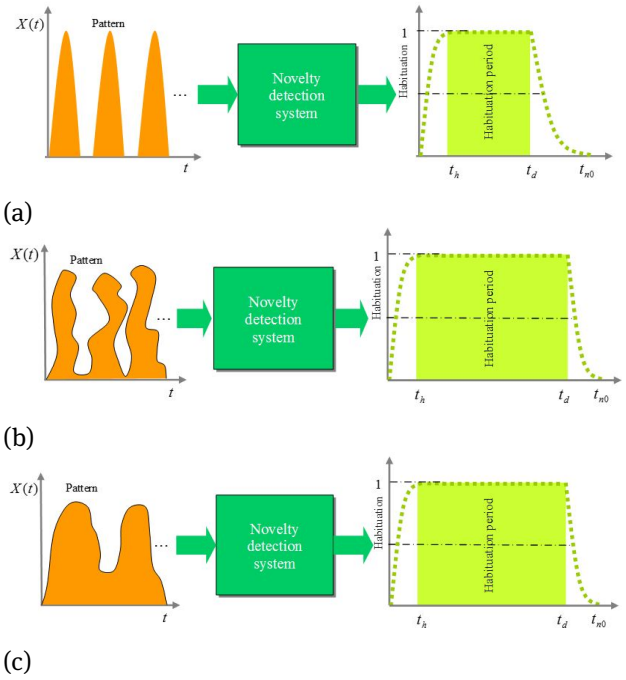

Figure 3: Habituation as a response to patterns of stimuli: (a) presentation of repeated patterns that generates standard habituation; (b)-(c) presentation of complex patterns prolongs habituation.

ers [11, 35, 36]. Most of the habituation based novelty detections are implemented using neural networks. Neto and Nehmzow [11, 35, 36] used Grow-When-Required (GWR) neural network for novelty detection where GWR uses the model of habituation defined by equation (1). In these implementations, visual features of the object captured from vision sensors require a number of geometric transformations.

3 Intrinsic Motivation

The concept of intrinsic motivation (IM) has been developed by psychologists to explain the behaviourist theory of the learning and drives [12–14]. Intrinsic motivational behaviour is observable in infants and young children that consistently try to grasp, throw, bite, squash or shout at new objects they encounter [37]. Such behaviours are also observable in adults. In the 1950s, psychologists started finding an explanation of IM and exploratory activities on the basis of the theory of drives [38]. Ryan and Deci [12] defined IM as doing an activity for its inherent satisfaction rather than doing it for reward or any other consequence. In other words, intrinsic motivations are drives for learning of skills, knowledge and compet-

ence [39]. Harlow [40] describes IM as the drive to manipulate and explore features. Festinger [41] defines IM as a minimisation process of cognitive dissonance between internal structures of cognition and perceived situation. Kagan [42] defines it as the reduction of uncertainty in the sense of the incompatibility between (two or more) cognitive structures, between cognitive structure and experience, or between structures and behaviour. Hunt [14] proposed the concept of optimal incongruity. Another group of researchers define IM as effectance [12], personal causation [43] and competence with self-determination [44]. A recently articulated and empirically well-supported theory developed by Redgrave and Gurney [45] shows that IM is related to a function called phasic dopamine (DA). DA activates the superior colliculus in the midbrain that generates a learning signal in the brain and which ceases after learning is complete. There are other views saying IM is a mechanism of measuring success in knowledge acquisition. Such mechanism drives animals to continue to engage in activities depending on their competence in achieving interesting outcomes, improvement in their capacity to predict, abstract, or recognise perceptions [46, 47]. Psychologists found evidence in animal and human that indicates IM play an important role in animals' behaviour that generates a reward for learning activities [12, 13, 44, 48].

There has been a growing interest among researchers in the cognitive computing, machine learning and robotics communities to model and reproduce IM that has been reported in the literature [37, 47–51]. A good insight of functions and mechanisms of IM and their computational models is reported in [37, 47]. An overview of a formal theory of IM and different ways of implementing basic computational principles is reported in [51]. Some researchers framed the problem of implementation in terms of reinforcement learning [37, 47, 50, 51]. The relation of IM with RL will be investigated further in the later section.

It is evident from the above discussion that the researches reported in the scientific literature hitherto are divided into three distinct streams based on the knowledge, competence and stimuli. In other words, IM can be categorised into three types: (i) Knowledge-based IM, (ii) Competence-based IM, and (iii) Novelty-based IM. These three types are discussed very briefly in the following section.

3.1 Knowledge-based IM

This computational approach is based on measures of dissonance between the situations experienced by an agent and the expectations about these situations of the agent [37, 47]. Motivation signal is generated based on internal prediction error between the agent's predictions of what is supposed to happen vs what actually happens when the agent executes a certain action. Most of the proposed models of IM are knowledge-based as they depend on the stimuli perceived by the learning system. A model proposed by Schmidhuber [52] consisted of adding to an RL agent an adaptive world model that learned to predict the next perception given the current perception and the planned action. The reward is the error in these predictions. In this way, the curious RL agent would be pushed to exploring poorly-known parts of the environment [47].

All the IM proposed by Berlyne [13] are knowledge-based in that they are related to the properties of the stimuli that the animal perceives and on their relation to the animal's knowledge. It is to be noted that Berlyne never made the distinction between knowledge and competence. Several similar knowledge-based proposals have been made in the psychological literature. These postulated either that animals are motivated to receive an optimal level of stimulation (e.g., [14, 53]) or that they are motivated to reduce the discrepancy (or incongruity or dissonance) between their knowledge and their current perception of the environment (e.g., [41, 42, 54]).

3.2 Competence-based IM

White [39] strongly advocated a competence-based view of IM, proposing that animals have a fundamental motivation for effectance, that is, for the capacity to have effective interactions with their environment. White's paper had a great influence on subsequent research on motivation. The important factors that make an activity intrinsically motivated are the sense of perception that the activity is self-determined, and the sense of competence, that is, the perception that the agent has (or is getting) mastery of the activity [39]. De Charms [43] proposed that personal causation, that is, the sense of having control over one's environment, was a fundamental driving force of human behaviour. Likewise, the theory of flow of Csikszentmihalyi [55] postulates that humans are motivated to engage in activities that represent an appropriate level of learning challenge i.e. the tasks that are familiar yet challenging, neither too easy nor too difficult to master given the individual's current level of competence.

The main difference between knowledge-based system and the competence-based system is that the knowledge-based system focus on agent's ability to model the environment and thus its capability to predict and expect what will happen next in that particular situation,

whereas the competence-based system focus on the skill improvement of the agent [47]. Knowledge-based applies measures that are related to the capacity of the system to model its environment whereas competence-based system applies the measures that are related to the system's ability to have specific effects on the environment.

3.3 Novelty-based IM

The most influential proposal was by Berlyne [13] suggesting that the factors underlying intrinsic motivational effects involve novelty, surprise, incongruity and complexity. He also hypothesized that moderate levels of novelty have the highest hedonic value because the rewarding effect of novelty is overtaken by an aversive effect as novelty increases. This is also consistent with many other research views holding that situations intermediate between complete familiarity (boredom) and complete unfamiliarity (confusion) have the most hedonic value. In a study of artificial creativity, Saundars and Gero [17] developed a system that demonstrates intrinsic motivation based on novelty and surprise following Berlyne's theories.

It is clear from the investigations in the preceding sections that a strong correlation exists between intrinsic motivation and novelty detection (ND) allied with the property of stimuli. In this case, novelty as such is almost tantamount to surprise, curiosity, incongruity, strangeness, or complexity. The inter-relationship between the two is illustrated in Figure 4.

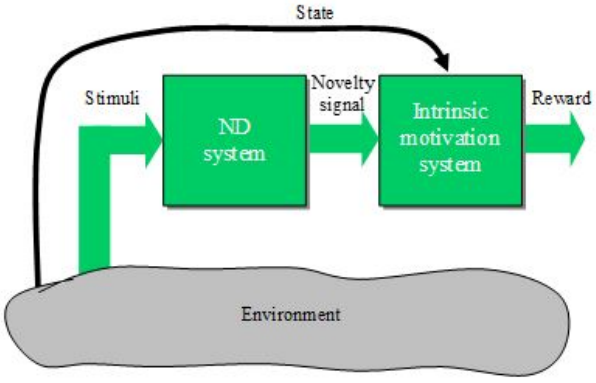

Figure 4: Functional dependency of ND and IM.

A significant amount of cognitive architectures and models of IM have already been developed in the literature [56–66]. There are several other models available in the literature in which the IM consists of some form of perceived novelty, prediction error, or learning progress of a world model [17, 62, 63]. A review of existing computational models of IM is available in [65]. Kaplan and Oudeyer [49, 60, 61, 64, 65, 67] have developed models of IM which mainly implemented different types of reward computation, which fits well with the relationship presented in Figure 4. Oudeyer and Kaplan [37] reported three broad classes of computational models of IM mainly based on information theory, knowledge-base, and competence-base. These models are well positioned in the computational RL framework. Thus, the typology relies on the formal description of the different types of reward computations that may be considered as defining an IM system. Sutton and Barto [68] provide a good account of possible RL algorithms in which the IM models can be plugged into those models proposed by Oudeyer and Kaplan [37] and applied to a range of applications.

The tendency to be intrinsically attracted by novelty has often been used as an example in the literature on intrinsic motivation. A simple way to computationally implement IM is to build a system that will generate a reward $r(e_k)$ based on the probability of an observed event $e_k \in E$, denoted as $P(e_k)$. The concept of entropy $H(E)$ over the set of events E can be used to characterize the shape of the distribution function:

$$H(E) = -\sum_{e_k \in E} P(e_k) \ln \left[P(e_k)\right]. \tag{2}$$

Oudeyer and Kaplan [37] provided a number of models of IM using the probability of observation of an event e_k. For uncertain motivation, the reward $r(e_k)$ for k-th event is inversely proportional to its probability of observation:

$$r(e_k) = C \left[1 - P(e_k)\right], \tag{3}$$

where C is a scaling parameter for adjusting reward values in a particular application. For local information gain motivation, the reward $r(e_k)$ is computed using the entropy between two-time instants t and $t + 1$:

$$r(e_k) = C \left[H_t(E) - H_{t+1}(E)\right]. \tag{4}$$

For distributional surprise motivation, the non-linear increase of reward $r(e_k)$ is computed as:

$$r(e_k) = C \frac{1 - P(e_k)}{P(e_k)}. \tag{5}$$

For distributional familiarity motivation, the reward $r(e_k)$ is proportional to its probability of observation:

$$r(e_k) = C\, P(e_k). \tag{6}$$

Various models of IM based on this simple mechanism have been implemented in the computational literature [12, 65, 69–71]. The reward computation mechanisms

described by the equations (3)-(6) can also be integrated within an RL architecture, which selects actions such that the expected cumulated sum of these rewards in the future will be maximized. These models are well suited to the novelty detection methods described in Section 2.

Using the relationship presented in Figure 4 and the reward mechanisms described by equations (2)-(6), a generic idealised reward function can be thought of which is proportionate to its novelty distribution. This is illustrated in Figure 5. Eventually, it is the model of IM incorporating the other state variables (both known and unknown) that will characterize the shape of the reward function. The event or object remains novel during the habituation period between $[t_h, t_d]$. The reward function starts earlier before it starts habituating, i.e. before time t_h. The reward function becomes zero when the novelty is habituated, i.e. after time t_d.

It is worth mentioning in the context of novelty and IM that some researchers also relate novelty with curiosity. In general, novelty and curiosity in IM are two different concepts. Gottlieb et al. [72] clearly differentiated between the two concepts. Novelty is a feature of an object whereas curiosity is a feature of the subject [16, 73]. Lowenstein [74] defined curiosity as information gap between what the observer knows and what he wants to know. Gottlieb et al. [72] proposed a curiosity-driven exploration learning model, which is different than the proposed framework in this paper.

4 Reinforcement Learning

Reinforcement learning is the process of improving performance through trial-and-error experience. The term RL comes from studies of animal learning in experimental psychology. The simplest RL algorithm is based on the common-sense idea that if an action is followed by an improvement or satisfaction in the state of affairs, then the tendency to produce that action is strengthened, that is, reinforced. This is known as Thorndike's [15] "law of effect."

In the standard RL framework, an agent is connected to its environment via perception and action. On each step of interaction, the agent receives an input $i \in s$ as some indication of the current state s of the environment. The agent then selects an action $a(t)$ to generate an output. The action changes the state of the environment, and the value of this state transition is communicated to the agent through a scalar reinforcement signal or reward $r(t)$. Figure 6 illustrates the mechanism of the standard reinforcement learning.

The key components of the RL algorithm are: Set of states $s \in S$, $S = \{s_1, s_2, \cdots, s_n\}$ and set of actions $a \in A$, $A = \{a_1, a_2, \cdots, a_n\}$, transition function $T\left(s, a, s^{'}\right) = p\left(s^{'}|s, a\right)$, i.e. the probability that an action a in state s will lead to the state $s^{'}$, a reward $r\left(s, a, s^{'}\right)$ received from the environment, and a policy π which determines what action to be taken in a particular state.

The goal of the agent is to maximize the cumulative reward also called return (value function or expected utility) in RL. It can learn to improve the performance over time by systematic trial and error and find an optimal policy. Some researchers, therefore, define RL as an optimisation problem. The value function is defined as:

$$V_{t+1}(s) = \max_a \sum_{s^{'}} T\left(s, a, s^{'}\right)\left[r\left(s, a, s^{'}\right) + \gamma V_t\left(s^{'}\right)\right], \tag{7}$$

where V is the value or expected utility at state s, T is the transition function, r is the reward, γ is the discount factor and t is a time step.

It is to be noted that reward in computational RL is different to the reward in psychology. What is known as a reward in RL is better called reward signal [76]. Reward acts as a positive reinforcement by increasing the frequency and the intensity of the behaviour that leads to achieving the goal [77].

A wide variety of RL algorithms are available in the literature. The RL paradigm divides neatly into search and learning aspects. The search for optimal policy has been taken by work in genetic algorithms and genetic programming, as well as some more novel search techniques [78]. In the learning aspect, there are two approaches: learning an optimal policy for an already known model [79–82] and learning an optimal policy when a model is not known in advance. There are two ways to learn policies for unknown models: model-based - where a policy is computed by learning a model and model-free - where a policy is computed without learning a model. Adaptive heuristic critic [83], TD(λ) [84] and Q-learning [85, 86] are widely used methods in model-free algorithm. Several researchers have used intrinsic motivation along with RL framework to form an autonomous learning agent [37, 48, 50, 51, 65]. There have been many researches reported in the literature [37, 50, 51, 65], where the problem of implementing intrinsic motivation has been framed in terms of RL. Sutton and Barto [68] provided a good account of possible RL algorithms which fits well with intrinsic motivation models.

The reward can also maintain learned behaviour by preventing extinction [87]. The rate of learning depends on the discrepancy between the actual and predicted re-

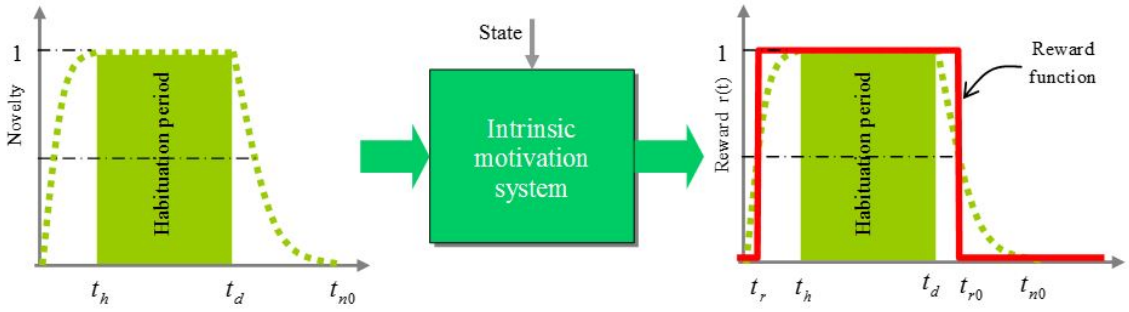

Figure 5: Reward as a function of the intrinsic motivation instigated by novelty.

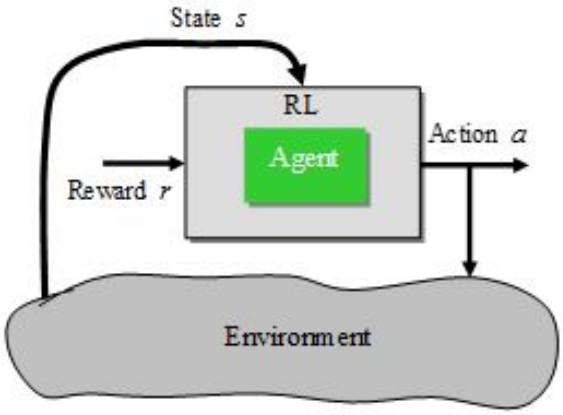

Figure 6: Framework of standard reinforcement learning [72].

ward [77, 88–92]. Rewards come in various physical forms and are highly variable in time. Reward information is extracted by the brain from a large variety of polysensory, inhomogeneous, and inconstant stimuli by using neuronal mechanisms [93]. One of the principal neuronal systems involved in processing reward information appears to be the dopamine system. An important finding reported by many researchers is the phasic activation of dopamine signal, which varies monotonically depending on the discrepancy between predicted and actual reward [94–96]. These researches suggest that dopamine response, denoted as D_{resp}, reports an error in the prediction of reward, i.e. D_{resp} can be expressed as

$$D_{resp} = r - \hat{r}, \tag{8}$$

where r is the reward occurred and $\hat{r}$ is the predicted reward.

Bromberg-Martin and Hikosaka [97] also reported the reward prediction error (RPE) mechanism found in a subpopulation of lateral habenula neurons in non-human primates. Ligaya et al. [98] showed how the anticipation of RPE affects agents and formulated the proposal in an RL model, which can account for a wide range of behavioural data. The objective of error-based learning is to reduce the average error to zero. A general problem of the error-based learning is that it cannot improve performance further when the average error reached zero. In this case, RL can guide the learning towards solution manifold [99]. The reward signal in RL provides less information than error-based learning, which results in slow learning. An alternative approach is the sensorimotor learning. Sensorimotor learning involves neural mapping between the motor and sensory variables, sensory-motor processing, sensory-motor transformations, and their modifications [100], which represent internal models. Significant advances have been made with applications to robotics [101, 102].

Singh et al. [76] used a linear reward function with two control parameters for adjusting the reward function. Oudeyer et al. [65] suggested integrating reward resulting from learning progress with other kinds of rewards, a weighted sum. A parameter $\lambda \in [0, 1]$ specifies the relative weight of each reward type defined as follows:

$$r(t) = \sum_i \lambda_i r_i(t - \tau) \tag{9}$$

where each $r_i(t)$ is referred to as a reward signal received at time t. τ is a time constant to adjust the reward function. The weights λ_i determine the contribution of these reward signals to the overall reward that the agent will learn to maximize throughout its lifetime. This formulation also corresponds to the RL algorithms provided by Sutton and Barto [68].

From the above discussion, the reward function $r(t)$ and the formulation of intrinsic motivation in terms of RL, a conclusion can be made on the relationship between intrinsic motivation and RL. The important thing here is to take care of the reward function of intrinsic motivation model and the reward function required by RL algorithm. The relationship is illustrated in Figure 7.

Some researchers suggested that shaping rewards do not alter the learning problem but they offer a possibil-

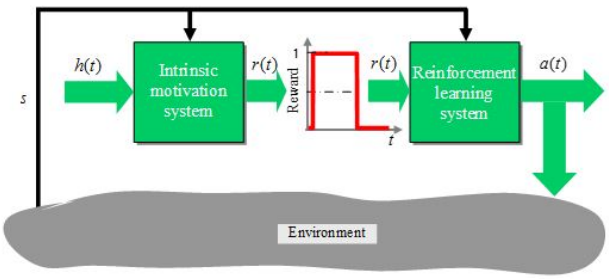

Figure 7: Learning as a function of reward and environmental states.

ity of providing a more informative performance feedback that can accelerate the learning [76]. Shaping by changing the reward signal can be done in two essentially different ways: permanently changing the signal and solving one new problem, or solving a sequence of problems leading to the original one. Ng et al. [103] have proved that the reward function r can be changed to $r + f$, while the original optimal policy is preserved. Shaping the reward function has been used by many researchers [104–106].

The study by Daddaoua et al. [107] with non-human primates combined with RL simulations also shows how reward predictive cues shape IM. Based on the investigations in sections 2, 3 and 4, we are proposing a workable integrated model incorporating ND, IM and RL that will be computationally viable for developing autonomous learning system. Such autonomous learning systems will find applications in developmental robotics.

5 Discussion on the integrated model of ND, IM and RL

In section 2, novelty detection methods have been discussed and it has been shown how the novelty of a pattern $X(t)$ can be translated into a habituation function $h(t)$ defined as

$$h(t) = \Phi\left(h_0, \tau, \alpha, X(t), w_s\right) \tag{10}$$

According to Eq. (1) in section 2, the habituation $h(t)$ can be defined as a function of initial habituation level h_0, time constants τ and α that control the habituation and the recovery rate respectively, and the weight vector w_s. In section 3, different models of intrinsic motivation have been discussed. In novelty-based intrinsic motivation, it has been shown how habituation function $h(t)$ can be utilised to compute the reward function. In principle, the reward function is defined as a function of habituation and state s as follows.

$$r(t) = \Gamma\left(h(t), s\right) \tag{11}$$

In section 4, different RL algorithms have been discussed with relation to intrinsic motivation. There are plenty of RL algorithms available in the literature, especially Sutton and Barto [68] provide a set of RL algorithms that suits well with intrinsic motivation models.

From the studies provided in Sections 2, 3 and 4, a three-way relationship can be established between novelty translated into a habituation function, IM translated into a reward function and RL system that produces an action. The compatibility of the models to each other and relationship between the three parts allow them to combine together as an integrated model. The integrated model is shown in Figure 8.

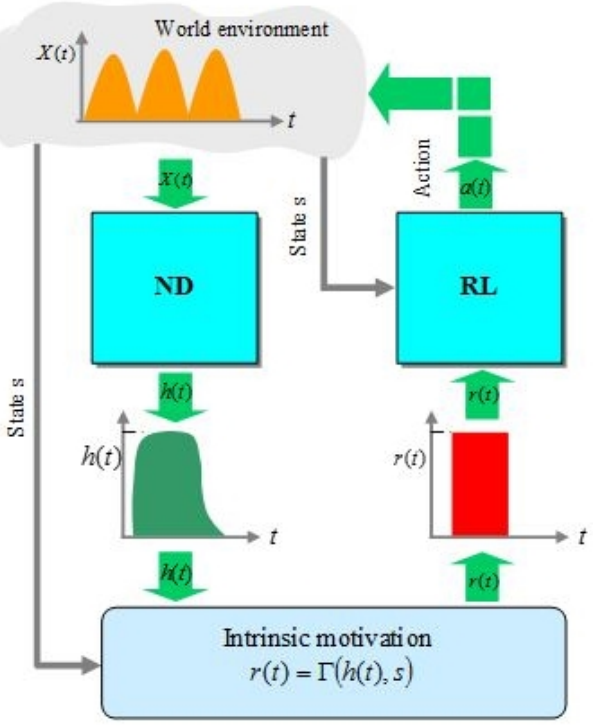

Figure 8: An Integrated model based on the tri-partite relationship between ND, IM and RL.

The computational models of ND, IM and RL reported in the scientific literature are reviewed and analysed in sections 2, 3, and 4 respectively. There have been many implementations of the individual models reported in the literature. The relationships between the three entities are suggestive to an integrated model that is computation-

ally viable, which can lead to design and development of autonomous learning systems.

6 Conclusion

Despite the diversity of the computational models of novelty, approaches of intrinsic motivation and RL, this paper presents a functional tri-partite relationship between the three. There is a point of convergence for all of them. Each of the described models defines a certain interpretation of the relationships between them in terms of properties of stimuli, novelty, habituation, intrinsic motivation, reward mechanism and RL. We believe that the relationship between the three entities ND, IM and RL can play a key role in design and development of autonomous learning systems and developmental robotics, an emerging area of robotics.

References

[1] M. Markou, S. Singh, Novelty detection: a review part 1: statistical approaches, Signal Processing, 83(12) (2003), 2481-2497

[2] S. Marsland, Novelty detection in learning systems, Neural Computing Surveys, 3 (2003), 157-195

[3] R. Saunders, J. S. Gero, The importance of being emergent, In: Proc. of the Conference on Artificial Intelligence in Design, 2000

[4] V. Chandola, A. Banerjee, V. Kumar, Outlier detection: A survey, Technical report, University of Minnesota, 2007

[5] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Computing Surveys, 41(3) (2009), 1-58

[6] V. Hodge, J. Austin, A survey of outlier detection methodologies. Artificial Intelligence Review, 22(2) (2004), 85-126

[7] M. Markou, S. Singh, Novelty detection: a review part 2: neural network based approaches, Signal Processing, 83(12) (2003), 2499-2521

[8] S. Marsland, U. Nehmzow, J. Shapiro, Vision-based environmental novelty detection on a mobile robot, In Proc. of the International Conference on Neural Information Processing (ICONIP'01), 2001

[9] S. Marsland, U. Nehmzow, J. Shapiro, On-line novelty detection for autonomous mobile robots, Robotics and Autonomous Systems, 51(2-3) (2005), 191-206

[10] H. V. Neto, U. Nehmzow, Automated exploration and inspection: Comparing two visual novelty detectors, International Journal of Advanced Robotic Systems, 2(4) (2005), 355-362

[11] H. V. Neto, U. Nehmzow, Incremental PCA: An alternative approach for novelty detection, In: Proc. Towards Autonomous Robotic Systems (TAROS'05), Imperial College, London, 12-14 September, 2005

[12] R. M. Ryan, E. L. Deci, Intrinsic and Extrinsic Motivations: Classic Definition and New Direction, Contemporary Educational Psychology, 25 (2000), 54-67

[13] D. E. Berlyne, Conflict, Arousal and Curiosity, McGraw-Hill, New York, 1960

[14] J. M. Hunt, Intrinsic Motivation and its Role in Psychological Development, Nebraska Symposium on Motivation, 13 (1965), 189-282

[15] E. L. Thorndike, Animal Intelligence, Hafner, Darien, 1911

[16] D. E. Berlyne, Curiosity and exploration, Science, 143 (1966), 25-33

[17] R. Saunders, J. S. Gero, Curious agents and situated design evaluations, In: J. Gero, F. Brazier (Eds.), Agents in Design 2002, Sydney, Key Centre of Design Computing and Cognition, University of Sydney, (2002), 133-149

[18] S. Wehmeier, Oxford Advanced Learner's Dictionary, edt., Oxford University Press, 7th edition, 2005

[19] 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 *Anticipatory Behavior in Adaptive Learning Systems*, Springer Berlin Heidelberg, (2009), 48-76

[20] F. E. Grubbs, Procedures for detecting outlying observations in samples, Technometrics, 11, (1969), 1-21

[21] E. Del Rosal, L. Alonso, R. Moreno, M. Vazquez, J. Santacreu, Simulation of habituation to simple and multiple stimuli, Behavioural Processes, 73 (2006), 272-277

[22] S. Sirois, D. Mareschal, Models of habituation in infancy, TRENDS in Cognitive Science, 6(7) (2002), 293-298

[23] P. M. Groves, R. F. Thompson, Habituation: a Dual-process Theory, Psychol. Rev., 77 (1970), 419-450

[24] R. F. Thompson, W. A. Spencer, Habituation: A model phenomenon for the study of neuronal substrates of behaviour, Psychological Review, 73(1) (1966), 16-43

[25] E. N. Sokolov, Higher nervous functions: The orienting reflex, Annual Review of Physiology, 25(1) (1963), 545-580

[26] J. C. Stanley, Computer simulation of a model of habituation, Nature, 261(5556) (1976),146-148

[27] N. K. Innis, J. E. R Staddon, What should comparative psychology compare?, International Journal of Computational Psychology, 2 (1989), 145-156

[28] D. L. Wang, A neural model of synaptic plasticity underlying short-term and long-term habituation, Adaptive Behavior, 2 (1994), 111-129

[29] D. Wang, M. A. Arbib, Modeling the dishabituation hierarchy: The role of the primordial hippocampus, Biological Cybernetics, 67 (1992), 535-544

[30] D. Wang, C. Hsu, SLONN: A simulation language for modeling of neural networks, Simulation, 55 (1990), 69-83

[31] M. A. Hunter, et al., Effects of stimulus complexity and familiarization time on infant preferences for novel and familiar stimuli, Dev. Psychol., 19 (1983), 338-352

[32] M. Velasco, F. Velasco, J. Machado, A. Olvera, Effects of Novelty, Habituation, Attention and Distraction on the Amplitude of the Various Components of the Somatic Evoked Responses, International Journal of Neuroscience, 5(3) (1973), 101-111

[33] C. Hutt, Degrees of Novelty and Their Effects on Children's Attention and Preference, British Journal of Psychology, 66(4) (1975), 487-492

[34] J. F. Ferreira, J. Davis, Attentional Mechanisms for Socially Interactive Robots – A Survey, IEEE Transaction on Autonomous Mental Development, 6(2) (2014), 110-125

[35] H. V. Neto, U. Nehmzow, Real-time automated visual inspection using mobile robots, Journal of Intelligent and Robotic Systems, 49(3) (2007), 293-307
[36] H. V. Neto, U. Nehmzow, Visual novelty detection with automatic scale selection, Robotics and Autonomous Systems, 55(9) (2007), 693-701
[37] P.-Y. Oudeyer, F. Kaplan, What is intrinsic motivation? A typology of computational approaches, Frontiers in Neurorobotics, 1(6) (2007), 1-14
[38] C. L. Hull, Principle of Behavior: An Introduction to Behavior Theory, Appleton-century-Croft, New York, 1943
[39] R. W. White, Motivation reconsidered: The concept of competence, Psychological Review, 66 (1959), 297-333
[40] H. Harlow, Learning and Satiation of Response in Intrinsically Motivated Complex Puzzle Performance by Monkeys, Journal of Comp. Physiol. Psychology, 43 (1950), 289-294
[41] L. Festinger, A Theory of Cognitive Dissonance, Stanford University Press, Stanford, 1957
[42] J. Kagan, Motives and development, J. Pers. Soc. Psychol., 22 (1972), 51-66
[43] R. De Charms, Personal Causation: The Internal Affective Determinants of Behavior Academic, New York, 1968
[44] E. L. Deci, R. M. Ryan, Intrinsic Motivation and Self-determination in Human Behavior, Plenum, New York, 1985
[45] P. Redgrave, K. Gurney, The short-latency dopamine signal: a role in discovering novel actions, Nature Review, Neuroscience, 7 (2006), 967-975
[46] G. Baldassarre, What are Intrinsic Motivations? A Biological Perspective, International Conference on Developmental Learning (ICDL-2011), 2011
[47] M. Mirolli, G. Baldassarre, Functions and mechanisms of intrinsic motivations: The knowledge vs. competence distinction, Chapter of book: G. Baldassarre, M. Mirolli (Eds.), Intrinsically Motivated Learning in Natural and Artificial Systems, (2012), 49-72
[48] E. Deci, Intrinsic Motivation, Plenum Press, New York, 1975
[49] F. Kaplan, P.-Y. Oudeyer, In search of the neural circuits of intrinsic motivation, Frontiers in Neuroscience, 1 (2007), 225-236
[50] M. Schembri, M. Mirolli, G. Baldassarre, Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot, In Y. Demiris, B. Scassellati, D. Mareschal (Eds.), The 6th IEEE International Conference on Development and Learning (ICDL2007), (2007), 282-287
[51] J. Schmidhuber, Formal Theory of Creativity & Intrinsic Motivation (1990-2010), IEEE Transaction on Autonomous Mental Development, (2010), 1-42
[52] J. Schmidhuber, Adaptive Confidence and Adaptive Curiosity, Technische Universitat Munchen, Technical Report FKI-149-91, 1991
[53] D. Hebb, Drives and the conceptual nervous system, Psychology Review, 62 (1955),243-254
[54] W. Dember, R. Earl, Analysis of exploratory, manipulatory and curiosity behaviors, Psychology Review, 64 (1957), 1-96
[55] M. Csikszentmihalyi, Flow: The Psychology of Optimal Experience, Harper Perennial, New York, 1991
[56] F. Kaplan, P.-Y. Oudeyer, Intrinsically Motivated Machines, M. Lungarella et al. (Eds.): 50 Years of AI, Festschrift, LNAI 4850, Springer-Verlag Berlin Heidelberg, (2007), 303–314
[57] A. G. Barto, O. Simsek, Intrinsically Motivation for Reinforcement Learning Systems, In: Proceedings of the Thirteenth Yale Workshop on Adaptive and Learning Systems, Yale University, New Haven, CT, (2005), 113-118
[58] A. Bonarini, A. Lazaric, M. Restelli, Self-development frame work for reinforcement learning agents, Proceedings of the Fifth International Conference on Development and Learning, Bloomington, IN, USA, 2006
[59] X. Huang, J. Weng, Novelty and Reinforcement Learning in the Value System of Developmental Robots, in C. G. Prince, Y. Demiris, Y. Marom, H. Kozima, C. Balkenius, (Eds.), Proceedings of the Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems 94, Edinburgh, Scotland, (2002), 47-55
[60] F. Kaplan, P.-Y. Oudeyer, Motivational principles for visual know-how development, In: Proceedings of the 3rd International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Lund University Cognitive Studies, Vol. 101, C. Prince, L. Berthouze, H. Kozima, D. Bullock, G. Stojanov, C. Balkenius (Eds.), Boston, USA, (2003) 73-80
[61] F. Kaplan, P.-Y. Oudeyer, The progress-drive hypothesis: an interpretation of early imitation, In Models and Mechanisms of Imitation and Social Learning: Behavioural, Social and Communication Dimensions, C. Nehaniv, K. Dautenhahn (Eds.), New York, Cambridge University Press, (2007), 361-377
[62] J. Marshall, D. Blank, L. Meeden, An emergent framework for self-motivation in developmental robotics, In Proceedings of the Third International Conference on Development and Learning (ICDL 2004), La Jolla, CA, (2004), 104-111
[63] K. Merrick, M. L. Maher, Motivated learning from interesting events: Adaptive, multitask learning agents for complex environments, Adaptive Behavior, 17(1) (2009), 7-27
[64] P.-Y. Oudeyer, F. Kaplan, V. V. Hafner, A. Whyte, The playground experiment: task-independent development of a curious robot. In Proceedings of the AAAI Spring Symposium on Developmental Robotics, 2005, D. Bank, L. Meeden (Eds.) (Stanford, AAAI), (2005), 42-47
[65] P.-Y. Oudeyer, F. Kaplan, V. V. Hafner, Intrinsic Motivation Systems for Autonomous Mental Development, IEEE Trans. Evolutionary Computation, 11(2) (2007), 265-286
[66] J. Schmidhuber, Curious Model-Building Control Systems, International Joint Conference on Neural Networks, 1991
[67] P.-Y. Oudeyer, F. Kaplan, Discovering communication, Connection Science, 18(2) (2006), 189-206
[68] R. Sutton, A. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998
[69] V. Fedorov, Theory of Optimal Experiment, Academic Press, New York, NY, 1972
[70] X. Huang, J. Weng, Motivational system for human–robot interaction, In: Proceedings of the ECCV International Workshop on Human-Computer Interaction, Prague, 2004
[71] N. Roy, A. McCallum, Towards optimal active learning through sampling estimation of error reduction, In Proceedings of the 18th International Conference on Machine Learning, Williamstown, MA, USA, Morgan Kaufmann Publishers Inc., 2001
[72] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, A. Baranes, Information-seeking, curiosity, and attention: computational and neural mechanisms, Trends in Cognitive Sciences, 17(11) (2013), 585-596

[73] C. Kidd, B. Y. Hayden, The psychology and neuroscience of curiosity, Neuron, 88(3) (2015), 449-460
[74] G. Lowenstein, The Psychology of Curiosity: A Review and Reinterpretation, Psychological Bulletin, 116 (1994), 75-98
[75] L. P. Kaelbling, M. L. Littman, A. W. Moore, Reinforcement learning: A Survey, Journal of Artificial Intelligence Research, 4 (1996), 237-285
[76] S. Singh, R. L. Lewis, A. G. Barto, J. Sorg, Intrinsically Motivated Reinforcement Learning: An Evolutionary Perspective, IEEE Transactions on Autonomous Mental Development, 2(2) (2010), 70-82
[77] W. Schultz, Multiple Reward Signals in the Brain, Nature Review: Neuroscience, 1 (2000), 199-207
[78] J. Schmidhuber, S. Heil, Sequential neural text compression, IEEE Transactions on Neural Networks, 7(1) (1996), 142-146
[79] R. Bellman, Dynamic Programming, Princeton University Press, Princeton, NJ, 1957
[80] D. P. Bertsekas, Dynamic Programming: Deterministic and Stochastic Models, Prentice-Hall, Englewood Cliffs, NJ, 1987
[81] R. A. Howard, Dynamic Programming and Markov Processes, The MIT Press, Cambridge, MA, 1960
[82] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, Inc., New York, NY, 1994
[83] A. G. Barto, R. S. Sutton, C. W. Anderson, Neuronlike adaptive elements that can solve difficult learning control problems, IEEE Transactions on Systems, Man, and Cybernetics, 13(5) (1983), 834-846
[84] R. S. Sutton, Learning to predict by the method of temporal differences. Machine Learning, 3(1) (1988), 9-44
[85] C. J. C. H. Watkins, Learning from Delayed Rewards. Ph.D. thesis, King's College, Cambridge, UK, 1989
[86] C. J. C. H. Watkins, P. Dayan, Q-learning, Machine Learning, 8(3) (1992), 279-292
[87] P. I. Pavlov, Conditioned Reflexes, Oxford University Press, 1927
[88] M. G. Baxter, E. A. Murray, The Amygdala and Reward, Nature reviews: Neuroscience, 3 (2002), 563-573
[89] W. Schultz, P. Dayan, P. R. Montague, A Neural Substrate of Prediction and Reward, Science, 275 (1997), 1593-1599
[90] W. Schultz, Getting Formal with Dopamine and Reward, Neuron, 36 (2002), 241-263
[91] P. N. Tobler, J. P. O'Doherty, R. J. Dolan, W. Schultz, Human Neural Learning Depends on Reward Prediction Errors in the Blocking Paradigm, Journal of Neurophysiology, 95 (2006), 301-310
[92] P. Waelti, A. Dickinson, W. Schultz, Dopamine responses comply with basic assumptions of formal learning theory, Nature, 412 (2001), 43-48
[93] W. Schultz, Predictive Reward Signal of Dopamine Neurons, Journal of Neurophysiology, 80 (1998), 1-27
[94] B. Brembs, F. D. Lorenzetti, F. D., Reyes, D. A. Baxter, J. H. Byrne, Operant Reward Learning in Aplysia: Neuronal Correlates and Mechanisms, Science, 296 (2002), 1706-1709
[95] C. D. Fiorillo, P. N. Tobler, W. Schultz, Discrete Coding of reward Probability and Uncertainty by Dopamine Neurons, Science, 299 (2003), 1898-1902
[96] J. R. Hollerman, W. Schultz, Dopamine Neurons Report an Error in the Temporal Prediction of Reward during Learning, Nature Neuroscience, 1(4) (1998), 304-309
[97] E. S. Bromberg-Martin, O. Hikosaka, Lateral habenula neurons signal errors in the prediction of reward information, Nature Neuroscience, 14(9) (2011), 1209-1216
[98] K. Ligaya, G. W. Story, Z. Kurth-Nelson, R. J. Dolan, P. Dayan, The Modulation of Savouring by Prediction Error and its Effects on Choice, eLIFE, 5, e13747, 2016
[99] D. M. Wolpert, J. Diedrichsen, J. R. Flanagan, Principles of sensorimotor learning, Nature reviews. Neuroscience, 12 (2011), 739-751
[100] D. M. Wolpert, J. R. Flanagan, Computations underlying sensorimotor learning, Current Opinion in Neurobiology, 37 (2016), 7-11
[101] H. Lalazar, E. Vaadia, Neural basis of sensorimotor learning: modifying internal models, Current Opinion in Neurobiology, 18 (2008), 1-9
[102] P. Lanillos, E. Dean-Leon, G. Cheng, Yielding Self-Perception in Robots Through Sensorimotor Contingencies, IEEE Transactions on Cognitive and Developmental Systems, 9(2) (2017), 100-112
[103] A. Y. Ng, D. Harada, S. Russell, Policy invariance under reward transformations: Theory and application to reward shaping, In Proceedings of the Sixteenth International Conference on Machine Learning, Morgan Kaufmann, (1999), 278-287
[104] V. Gullapalli, Reinforcement Learning and Its Application to Control, PhD Thesis, University of Massachusetts, COINS Technical Report 92-10, 1992
[105] M. J. Mataric, Reward functions for accelerated learning, In: W. W. Cohen, H. Hirsh (Eds.), Machine Learning: Proceedings of the Eleventh International Conference, Morgan Kaufmann, CA, 1994
[106] J. Randløv, P. Alstrøm, Learning to drive a bicycle using reinforcement learning and shaping, In: Proceedings of the Fifteenth International Conference on Machine Learning, Morgan Kaufmann, (1998), 463-471
[107] N. Daddaoua, M. Lopes, J. Gottlieb, Intrinsically motivated oculomotor exploration guided by uncertainty reduction and conditioned reinforcement in non-human primates, Nature Scientific Reports, 6(20202) (2016)