Artificial Motivations based on Drive-Reduction Theory in Self-Referential Model-Building Control Systems

Moritz Schneider, Jürgen Adamy
Institute of Automatic Control and Mechatronics
Laboratory of Control Methods and Robotics
Technische Universität Darmstadt
64283 Darmstadt, Germany
{schneider,jadamy}@rmr.tu-darmstadt.de

***Abstract*—Motivation and emotion are inseparable component factors of value systems in living beings, which enable them to act purposefully in a partially unknown and sometimes unforgiving environment. Value systems that drive innate reinforcement learning mechanisms have been identified as key factors in self-directed control and autonomous development towards higher intelligence and seem crucial in the development of a concept of "self" in sentient beings [1]. This contribution is concerned with the relationship between artificial learning control systems and innate value systems. In particular, we adapt the state-of-the-art model of motivational processes based on reduction of generalized drives towards higher flexibility, expressivity and representation capability. A framework for modelling self-adaptive value systems, which develop autonomously starting from an inherited (or designed) innate representation, within a learning control system architecture is formulated. We discuss the relationship of anticipated effects in this control architecture with psychological theory on motivations and contrast our framework with related approaches.**

I. Introduction

Most technical systems are based on the principle of direct manipulation by a (human) operator, where it is crucial that the operator has (1) a powerful interface to the inner workings of the artifact and (2) detailed knowledge of the artifact and its relation to the outside world, in order to keep the artifact functional and achieving its purpose. Intelligent systems, on the other hand, should be able to achieve their goals autonomously while staying intact and respecting important environmental constraints, with a minimum of human intervention.

Learning control systems are artificial systems exhibiting some form of control over another system, the environment, as to fulfill certain criteria defined w.r.t. to controller/environment interaction, using innate or even self-generated optimization algorithms but with none or only little prior knowledge of their environment. The system starts with some basic structure, e.g. an innate optimization method or an architecture of learning components, which it may continually modify, aiming at the optimization of some (usually externally defined) criterion over the course of its existence.

Consider a robot designed to explore the surface of a distant planet. The robot must be able to cope with adversaries on small timescales all by itself, since the long signal running time to earth might be prohibitive for full telecontrol. Also, action planning respecting available resources has to be done in relative autonomy when availability and needfulness of these resources may change so fast that pure telecontrol is again unfeasible. When considering multiple conflicting goals, the robot must be able to decide which one to pursue at which instant of time as to sustain its own reliable functioning, and, if it continually needs to process and learn before unknown causal relationships between sensory signals, it also must consider when to spend time or another important resource for learning something new and if that particular something is worth the effort. This implies complex representations of and cognition about subjective value w.r.t. the current understanding of the state of the environment compared to perceived own internal states. A control system which adapts its own internal processes towards enhancing its capabilities to represent and anticipate determinants of causal forces in its interaction with the environment, which are relevant to its basic innate value system, is called a self-referential learning control system [2]. In [1], a sketch for a cognitive architecture based on this principle is given by rigorous consideration of structural design requirements and capabilities that must be respected when implementing such architectures. In this contribution, we are concerned with their notion of value systems and show how an adaptive architecture of artificial motivational states may serve as a conceptual basis for design and implementation of systems of innate and derived subjective value.

Over the course of evolution, many higher animals have developed internal mechanisms that are strikingly successful at solving the type of problem discussed above. Different representations of subjective value are hard-coded into biological nervous systems through the functional interactions between different neurotransmitters with each other and the chemical structure and physical behavior of neurons [3]. It is meanwhile well established that e.g. phasic Dopamine neuron activity is a key factor in the representation of achievement motivation, wanting, and affective arousal in biological systems [4]. The interesting point is how a continually self-improving control system can represent, manipulate and enhance its internal representations of different modalities of value and also may create subjectively consistent complex amalgams mediating responses to and cognitions about complex stimuli and cognitive representations. The authors of [2] argue that these processes may play a pivotal role in the self-organized creation

978-1-4799-1959-8/15/$31.00

of cognitive semantics and a sense of self in human beings. Building artificially motivational systems is thus not only a biologically inspired approach to make artificial intelligent systems more adaptive to their environment, but may be also a constructive way of explaining the development of intelligent purposive cognition and behavior [1], [5]. (Note, however, that this does not require normative models to align perfectly with neurophysiological reality in order to capure the mechanisms of motivational processes [5].) This issue is closely tied with artificial emotions which enable further situational adaptivity of the system to its environment under constraints given by its knowledge about the interactions of itself with the environment [6]. In order to account for this issue in future work, we are explicitely concerned with motivational processes within control architectures that build models of the interaction with their environment and use these models to reason about their current and future motivational states and the uncertainty they associate with it, whereas most work towards artificially motivated learning systems has been formulated explicitely in the context of a type of model-free reinforcement learning.

This paper is organized as follows: After a brief overview on approaches to artificial motivations and their use in reinforcement learning agents, a general formal framework for elicitation, representation, and modulation of artificial motivational states is presented, building upon work by [1], [7], [8], [9], [10]. In particular, we highlight important implications of interaction between cognition, e.g. perception, memory, action selection and planning, and incorporation and use of complex models of subjective value into intelligent control systems. The last section is concerned with self-adaptive motivational systems, an idea first explored by [11] and subsequently refined in [10].

While being theoretical in nature, in this article also some implementation issues and possible learning methods are discussed towards the realization of self-referential learning and self-optimizing control systems.

II. Artificial Motivations: Theory and Formal Framework

In the beginning of artificial intelligence research, minds were conceived purely as problem solving devices [12], without further specification of why they solve particular problems while apparently completely ignoring others. Today, a large body of theories on motivations and their modulating and directing influence at different stages of information-processing in humans is available that tries to tackle the problem from normative, descriptionalist, neurophysiological, or constructivist viewpoints. On the grounds of these findings, in the last 25 years there has been a growing interest of integrating plausible motivational mechanisms in artificial systems, towards further understanding of biologically implemented minds, widening the scope of AI research, or construct intelligent artifacts with increased autonomy. While a comprehensive review of different lines of research is widely out of the scope of this paper, we would like to begin with a brief discussion of basic theoretical concepts of the field and different approaches to the construction of artificially motivated intelligent systems.

A. Extrinsic vs. Intrinsic Motivations

In classical behaviorist psychology, the prevailing notion for describing why higher organisms behave in a certain way has been the notion of (basic) drives. It essentially describes the ends of behaviors as reduction of different categories of deprivations. For example, Hull's famous model postulated hunger, thirst, sex and avoidance of pain as basic drives [13]. When cognitivism became the leading paradigm in academic psychology in the 1960s, it became apparent that basic physiological drives do not suffice to explain the wide variety and complexity of human behaviors. Basic drives - subsumed under the notion of extrinsic motivations - became conceptually separate from intrinsic motivations (IMs), such as curiosity, autonomy, or competence. [14] notes that the conceptual distinction (and hence classification of IMs as non-drives) was based on an assumed physiological distinction that turned out to be false. Note, that the physiological implementation in natural systems as humans differs: whereas lower level homeostatic drives are explicitely represented in neural structures, IMs seem to appear in an emergent fashion. However, issues concerning biological implementation are not necessarily important for abstract models aiming at artificial systems [5].

Based on a unified view of drives and IMs, [14] suggested a model of 16 basic desires encompassing basic drives as well as IMs. [9] describes the motivational subsystem of the cognitive architecture CLARION where a related configuration of basic desires drives behavioral and learning processes within the architecture. This model encompasses physiological, as well as a quite high number of high level cognitive and social motivations which have been found to be relatively independent from another. [15] followed a similar route in the conception of the motivational system within the cognitive architecture Psi [6]. In Psi, several subcategories of the main demand categories (1) physical, (2) cognitive and (3) social needs are distinguished. Psi's demand system works as a simple linear integrator system, which means that the resulting demand systems resembles a tank model, whose level is determined by in- and outflowing current. Physical demands are similar to Hullian drives, cognitive demands point at the needs of a complex model-building control architecture, e.g. competence or certainty w.r.t. the system's environment, and the third considers a Psi agent as an entity among others, having needs for affiliation or status. A comprehensive overview on Psi in English language is presented in [6] and a self contained description of the motivational subsystem of the architecture can be found in [7].

Coming from a different route, a number of artificial intelligence researchers have formulated models of artificial intrinsic motivations. In the reinforcement learning community, for example the exploration vs. exploitation bias has been a long-standing issue: If an agent explores its possibilities for action in the world to the end of constructing behavioral policies (which are in some sense optimal) with the aim of exploiting these for maximizing its external reward, it is not trivial when the agent should do what it knows is good and when it should rather search for better strategies than the ones which it already knows. Two prominent approaches in this fashion are artificial curiosity [16] and intrinsically motivated reinforcement learning [17].

Artificial curiosity (in its most simple formulation, as

originally suggested in [18]), in an agent that interacts with an environment and in the course of this interaction builds models of its environment, means that the agent trains its action policy to supply it with new, i.e. unexpected or unpredicted, perceptual input which appears regular enough to be learnable. The agent should therefore be capable of (1) finding good approximations of the state/action values faster than a purely probabilistic exploration scheme and (2) gather information on regularities within the environment without explicitely relating the process to the material outcome of interacting with them. A more detailed treatment of this mechanism can be found in [16] and numerous other works on this topic by the same author. An interesting variant of this idea is explored in [19], where the concept of artificial curiosity is explored from a developmental perspective. Therein, learning progress is defined as a generalized drive and used to bias a modular structure of reinforcement learners towards development of complex representations and behaviors.

In context of the latter, the first approach to self-modifying motivational systems has been made by [11]. In the framework of intrinsically motivated reinforcement learning, the authors propose an additive combination of basic reward, the IM component, and other functions of the perceptual state of the agent which were evolved using a genetic algorithm. These other possible reward functions were kept when they proved useful for the expected total basic reward which the system has achieved. The idea of using additional virtual reward to speed up learning to collect basic reward is, however, not new and is often called reward shaping in reinforcement learning research, albeit these approaches are generally not directed at recreating psychological mechanisms, at least not in an explicit way.

B. Motivations as Generalized Drives

Hull described motivational states as sensed deprivations w.r.t. physiological states, which drive behavior towards maintaining a homeostatic balance. Later, his model was used to describe also intrinsic motivations, i.e. motivational states that do not refer to physiological state, like curiosity or fairness as differences between some desired state and the actual state of some reference variable, resulting in what has been called "generalized drives".

Virtually all models of artificial motivations use reduction of generalized drives to create meaningful feedback signals for the agent about the quality of its behavior. There are, however, differences in the definition of the urge represented by the distance between current and desired value. The line of research followed by [15], [6], [7], [9] considers a single target value and [8] a target region bounded from below by a given threshold. On the contrary, relevant observed quantities in real world processes are often required to obey upper as well as lower constraints, as e.g. body temperature in most lifeforms. It is thus reasonable to account for this issue in a general framework.

More formally, if we write the agent as a dynamical system

$$\begin{aligned} x^A(k+1) &= f(u^A(k), x^A(k)), \\ y^A(k) &= g^A(u^A(k), x^A(k)), \end{aligned} \tag{1}$$

with $u^A \in U \subseteq \mathbb{R}^m$ denoting the sensory input to the agent, $x^A \in X \subseteq \mathbb{R}^n$ the agent's internal states, and $y^A \in Y \subseteq \mathbb{R}^p$ its action on the environment, we can designate a mapping of the state-space of the agent X^A to the space of motivationally relevant states $x^{M(A)} \subseteq \mathbb{R}^q$, which we will only refer to implicitly as Ψ (note that in this formulation, we do not require $q \leq n$, meaning that any state of the agent can have more than one corresponding motivational states, which allows e.g. for the definition of motivational states which do not have an unimodally bounded desirable region). This means that there can be more than one critical point marking the boundary of a desirable region, even for the same state variable. Since the partial transformation Ψ may also invert the orientation, we can have a desirable region bounded from above as well as from below, for the same state. Additionally, Ψ can be used to formulate a demand such that the corresponding motivation implements an avoidance motivation w.r.t. a certain region.

Now we can characterize a demand d_i as a mapping

$$\begin{aligned} &d_i : X_i^{M(A)} \times X_i^{M(A)} \to \mathbb{R}, \\ &x_i^{M(A)}, s_i \mapsto \mathbb{D}(x_i^{M(A)}, s_i), \end{aligned} \tag{2}$$

taking the i-th element of the motivational state vector $x^{M(A)}$ and the critical point s_i defined on $x_i^{M(A)}$ to the corresponding demand strength using $\mathbb{D}$, which is usually some generalized distance measure. Note, that s_i does not need to be fixed and that for some motivationally relevant state variables it makes sense to consider them as being directly controlled by the agent, as will be seen later.

Then, we can use $d_i,\ i \in \{1, \ldots, q\}$ as a measure for the urge or desire associated with the motivational variable. In our case, we define the magnitude of the desire associated with a motivationally relevant state variable $x_i^{M(A)}$ at time step k as

$$d_i(k) = (\max(0, s_i(k) - x_i^{M(A)}(k))^m)^{\frac{1}{m}} \tag{3}$$

for $m = 1, \ldots, \infty$. Therein m is a design parameter that controls the small-scale behavior of $d_i^*(k)$, e.g. $m = 1$ implies proportionality between target value deviation and motivation strength, while $m = 2$ enforces comparatively smaller motivation strengths for smaller target value deviations. $m = 2$ seems like a reasonable default value. The motivational state vector of the agent $D^A(k)$ can then be written as

$$D^A(k) = (d_1(k), \ldots, d_q(k)). \tag{4}$$

Up until now, this is quite the same procedure as in [8] or [7], with the difference that we allow for different (possibly upper as well as lower) bounds on the motivationally relevant variables, whereas [8] considers single target points and non-oriented deviations, and the setting in [7] allows only for specification of lower bounds. However, the definition of d_i is formally identical to the variant introduced by [8].

Defining motivations corresponding to homeostatic regulation of internal states, such as hunger or thirst, is straightforward in this setting, but is this also the case for "growth"-related motivations, such as curiosity or autonomy? The answer is yes, motivations related to growth may be modeled as drive reduction when the sought gradient of growth is represented through the dynamics of the corresponding set-points s_i. In the case of curiosity, for example self-monitoring

and -modeling in closed loop with the environment would enable the agent to estimate about which states it has to gain knowledge in order to be able to achieve a certain goal and let the corresponding drive strength be proportional to the importance of said goal. A simpler variant of set-point control would be proportionally increasing the lower set-point of a state representing information gain every time this drive is reduced. An interesting feature of this variant is that the agent's performance in regard of this motivational state will subsequently influence motivation strength. In short, the agent will be less motivated to be curious, when according to its experience, curious exploration does not lead to any knowledge increase, an issue also reflected in the evolution of the notion of artificial curiosity [16].

It is our goal to define reward in terms of subjective value of the agent corresponding to different situations, thus the employed formulation should be as flexible as possible. In particular, it would seem unrealistic that the demand magnitudes for different motivationally relevant states only depends on the magnitude of the target value deviation. In particular, there may be a given importance ordering relation on different motivational states (e.g. avoidance of physical damage may be more imporant than approach of a task-oriented approach motivation). On the other hand, the relative importance of different motives may also depend on the magnitude of other motivational states, i.e. the need for avoidance of physical damage might suppress the subjective need for energy intake for a while, even if failure to achieve one of the associated goals might result in a fatal result. However, spending only a small amount of the energy resources left while fleeing a predator is less likely to end successful than devoting almost all available energy to the escape and caring for the energy problem as soon as it is safely possible.

Default relative importance and adaptive changes in relative importance through mutual inhibition of activation of different motivational states can be implemented by the following simple extension of (3) and (4):

$$\begin{aligned} d_i^*(k) &= C_i^T \cdot D^A(k), \\ &= \sum_{j=1}^{q} c_{ij} d_j(k), \\ D^{A,*}(k) &= C \cdot D^A(k), \end{aligned} \tag{5}$$

where $C \in \mathbb{R}^{q \times q}$ is a matrix called drive controller and C_i denotes the i-th row of the drive controller matrix. C can be set in a variety of fashions (where I denotes the identity matrix of dimension q):

- $C = cI$, $c \in \mathbb{R}$ regulates the sensitivity of urge elicitation in the face of deviations of demands for their target values.
- $C = c^T I$, $c \in \mathbb{R}^q$ implements different priorities for different demands.
- a general $C \in \mathbb{R}^{q \times q}$ allows for different urges to interact directly with each other. The interaction can be excitatory or inhibitory.
- in any variant, employing a time-dependent or adaptive $C(k)$ allows for a greater level of cognitive control over the motivational system.

So far we have specified the motivationally relevant substate of the agent state at some given time, defined the desired region for each motivationally relevant variable by a compact interval and derieved the strength of the corresponding desire.

This framework extends approaches by [15], [6], [8], [11] and allows for an integrated treatment of upper and lower constraints, respectively, which is not possible in the framework of [8], due to the consideration of simple setpoints. Our framework therefore generalizes the above mentioned settings, since simple setpoints can be emulated with very short intervals.

Moreover, the drive controller formalism goes beyond static weighting of different demand strengths by predefined preference factors, which already can serve as powerful models for personality traits, as elaborated in context of the "Big Five" personality model in [6], [9], to cognitive modulation of motivational processes. In context of model-free RL it is noted by [8] that mutual inhibition of drives is not necessary since stochastic selection already may account for inhibition in habitual action selection. However, later on we will intertwine model-free and model-based RL to lay out a system which can produce habitual responses as well as goal oriented action planning driven by cognition. In the latter setting, the drive controller provides a formal basis for making explicit tradeoffs between the agent's goals and serves as a convenient and structured representation of dynamic preferences to enable external as well as internal coping and self-adjustment strategies in the face of difficult problems.

Next, we formulate the generation of pleasure and displeasure signals from a collection of urge values.

C. Value-Signals Generated from Drive Reduction

In the following, we derive two valency signals per motivational state, which can be combined to a single, global evaluation signal, e.g. by taking a convex combination, but also may be available separately to different components of the cognitive architecture of the agent, similar to what has been observed in humans [3].

Since the motivational state space results from a transformation of the whole state space of the agent, a trajectory of agent states starting at some time step k through time steps $k+1, \ldots, k+n$ for some $n \geq 1$, i.e. $x^A(k), \ldots, x^A(k+n)$ should result in a corresponding trajectory through motivational state space $x^{M(A)}(k), \ldots, x^{M(A)}(k+n)$ with associated demand vector dynamics $D^{A,*}(k), \ldots, D^{A,*}(k+n)$, where each $x^A(k+i)$ must obey

$$x^A(k+1) = f^A(u^*(k), x^A(k)), \tag{6}$$

for some environmental action $u^*(k)$. The corresponding motivational state transition from $x^{M(A)}(k)$ to $x^{M(A)}(k+1)$, which is characterized by

$$x^{M(A)}(k+1) = \Psi(f^A(x^A(u^*(k), x^A(k)))) \tag{7}$$

then yields successive demand states $D^{A,*}(k), D^{A,*}(k+1)$.

Drive reduction theory states that reducing an urge results in a reward signal proportional to the magnitude of the drive reduction and that an increase in urge magnitude results in a proportional displeasure signal. For every component $x_p^{M(A)}$

of the motivational state vector, we may define the pleasure V^+ and displeasure V^- resulting from the transition between $x^A(i)$ to $x^A(i+1)$ as

$$\begin{aligned} V_p^+(i+1) &= \max(D_p^{A,*}(i) - D_p^{A,*}(i+1), 0), \\ V_p^-(i+1) &= \max(-(D_p^{A,*}(i) - D_p^{A,*}(i+1)), 0), \end{aligned} \tag{8}$$

and calculate the global valency state by summing the components

$$\begin{aligned} V^+(i+1) &= \sum_{j=1}^{q} V_j^+(i+1), \\ V^-(i+1) &= \sum_{j=1}^{q} V_j^-(i+1). \end{aligned} \tag{9}$$

Finally, the reward obtained by the control system at time step $i+1$ may be defined as

$$r(i+1) = \alpha V^+(i+1) - \beta V^-(i+1), \tag{10}$$

where α, β are adaptive parameters that, similar to the drive controller, may reflect static/innate preferences or personality traits of the control agent, or might depend on current situational evaluations. This is motivated by the observation made in psychological research on emotions that individuals in different emotional states seem to evaluate and weight negative and positive prospects differently. For instance, individuals in sadness-like states tend to overestimate severity of negative outcomes or obstacles while happy individuals systematically underestimate them compared to anticipated positive outcomes. For instance, studies have found that hearing sad music while standing at the bottom of a hill influenced perception of the steepness of the hill in participants of the experiment. [20] describe this phenomenon as "emotions can make mountains out of molehills". This effect provides self-driven situational adaptivity: If some situation repeatedly causes sadness, underestimating one's own coping potential for this (type of) situation may prevent one from wasting resources and eventually withdraw completely from such seemingly desperate situations. If the sadness state is somewhere represented within the agent's state variables, it may in turn also be used as an aversive motivation.

Formally, our definition of the reward function is only a mild generalization of the formulation presented in [8], therefore our model inherits some interesting traits from this formulation, which the authors have used to demonstrate the behavioral plausibility of their model.

For the following discussion, assume α, β are fixed at some $c_1, c_2 \in \mathbb{R} \setminus \{0\}$.

- $\frac{dr}{dV^+} > 0, \frac{dr}{dV^-} < 0$, i.e. the reward is an increasing function of the magnitude of pleasure and simultaneously decreasing in the magnitude of displeasure
- $\frac{dr}{d|s_i - D_i^{A,*}|} > 0, i \in \{1, \ldots, q\}$ and $D_i^{A,*} < s_i$, which means that the rewarding value of a stimulus changing a demand level increases with increased deprivation of this demand level

Mutual excitatory and inhibitory influences of demands which carry different levels of subjective importance is in our framework already encompassed with the drive controller formalism.

D. Self-Referential Control through Motivated Approximate Dynamic Programming

Approaches for self-referential control through artificial motivations are often used with a form with an associative learning mechanism to enable possibilities for learning and self-improvement. Psi-theory accomplished this through a cognitive architecture built upon Hebbian learning mechanisms, while [8], [9] formulate the problem of action selection in a standard reinforcement learning framework. In particular, [8] discusses the connections of different indirect or direct RL approaches to self-referential control in a quite detailed way and relates learning theory with results on traits of biological motivational systems and motivated behavior in animals. In this section, we dive further into this issue with focus on the possible implications of different systems utilizing model-free in contrast to model-based RL or even both at the same time, with emphasis on cognitive processes acting in each variant and their relationship to psychological theories of motivation and emotion.

1) Reinforcement Learning and Approximate Dynamic Programming with and without a World Model: The following part is not meant as an introduction to reinforcement learning but just to recapture the formal setting and discuss some issues related to the difference between direct (model-free) and indirect (based on a world-model) reinforcement learning.

In a setting where an agent A interacts (in discrete time) with an environment E which, at some points, gives a reward r to the agent as an evaluation of its performance. At each time step, upon receiving a perceptual (and possibly a reward) signal from the environment, the agent chooses an action which of course may influence future reward signals. In deterministic or stochastic settings, an agent with a Markovian interface to its environment may solve this problem using Bellman's optimal value (not to be confused with the concept of a value system) function

$$V^*(x) = \max_a (r(x) + \gamma V(T(x,a)), \tag{11}$$

where $T(\cdot)$ is the (usually not known) transition function of the environment and $\gamma \in (0,1)$ is a discount factor. In practice, this is often done by actually considering a state-action value function $Q(s,a)$, since then it is convenient to just choose an action maximizing $Q(s_0, \cdot)$ for any given initial state s_0. Often, a parametric function approximator is employed to represent V, Q or even the action selection policy $\pi : X \to A$.

The idea behind dynamic programming value functions is to enable optimal decisions using only one-step ahead predictions, which is also why standard methods that solve this problem, such as temporal difference learning of the value function, usually require the interface between agent and environment to be a Markov decision process (MDP). It is, however, far from true that RL is only possible in MDPs: besides using direct policy search techniques which do not rely on Dynamic Programming, the problem can for example be alltogether circumvented when the value function of a state x_i may take past observations $x_{i-1}, \ldots, x_{i-n}$ into account. For instance, [21] studies the use of recurrent neural networks in reinforcement learning within non-Markovian settings.

It might also become difficult for temporal difference learning algorithms to approximate V^* or Q^* in very large

Markovian environments, since approximating value functions requires that the learning algorithm can observe the long-term effects of different action choices a in the same state s, meaning that large problem spaces can be quite difficult to sample, depending on the complexity of the true value function of the environment w.r.t. the defined reward function. On the other hand, if a world model $T : X \times A \to X$ is available, or if there can be constructed at least an approximation to it using past experience, we can construct a value function V, action-value function Q and even behavioral policy $\pi : X \to A$ by using internal simulations, e.g. rollout planning algorithms. While the creation and maintenance of such a world model imposes its own cognitive workload and subsequently some difficult problems, it also allows for explicit reasoning with and on uncertainty.

2) Reinforcement Learning and Artificial Motivations: In Sect. II.B., we have basically defined an artificial motivational system as an internal reward signal generator, whose activity depends on internal states or perceptions of the agent. This can be directly incorporated into a RL framework, just by substituting (or enhancing, e.g. by additive combination) environmental reward with the reward signal generated by the motivational system, as suggested by [8]. Therein, also an interesting consequence of action selection based on model-free in contrast to using model-based RL is discussed. The authors argue that choosing actions based on iterated value functions (aggregated experience from many similar but in details different situations) works as a habitual system and how it may perform predictive homeostatic control (actions selection at time step k in some state depends on the value function at that state which encompasses discounted rewards from situations that usually occur after the current). Model-based RL is envisioned to incorporate more aspects of explicitely goal-directed behavior.

Model-based RL, however, offers the further possibility of framing search for criterions for good behavior, i.e. identifying additional demands from striving to fulfill current demands, in exactly the same framework as the search for a behavioral policy. This is because an architecture of components determining a behavioral policy by an optimization method applied to an existing world model can serve as a model of the control agent itself, which it can use to evaluate actual or simulated changes to the whole architecture or single components.

Moreover, using environmental models it is possible to evaluate and compare different representations of value or behavioral policies in parallel. Traces of motivational signals can be saved in a memory-like structure, along a history of internal and external state values, and serve subsequently for instantiation of internal simulations, derivation of different types of error signals (for reward intake, reward quantities, predictions, etc.) for the information processed by the agent to determine what may currently be interesting This knowledge can be further exploited when using the world model to plan value representations or controller models by providing a bias towards interesting or relevant situations in a sampling-based planning scheme, instead of simple Monte-Carlo sampling. The authors of [2] suggest that solving the problem of building and maintaining an adequate model of environmental reactions to one's own actions to enable and sustain subjectively consistent self-directed purposive behavior w.r.t. an innate value system affords quite interesting and important mental mechanisms living at the borderline of the traditional distinction between affect and cognition within psychological literature. Furthermore, following the basic ideas of [18] and their normative implications for indirect reinforcement learning systems, it is quite possible that some well-known motivational states, such as curiosity, in biological lifeforms actually have their causal precursors in the fundamental problems that arise when trying to build a useful (w.r.t. a given task) model by observation through direct experimentation. Since a controller utilizing a model it needs to build by itself from its own experience inherently has a dual objective task to perform, a system building and acting on an explicit representation of how much it needs to know about which states to be able to achieve its task has advantages over randomly exploring controllers when the environment is regular enough for the first option to be feasible.

Eventually, we would like to highlight some interesting features of a dual representation of value. Let the agent A and its environment be defined as before. Assume the agent has obtained, for simplicity, a single world model represented by some parametric approximator $\mathbb{M} : X \times A \to X \times \mathbb{R}$, e.g. a neural network architecture of arbitrary size and complexity which outputs e.g. mean and variance of a Gaussian distribution of possible states s_{i+1} after performing action a in state s_i and further a collection of controller modules $\mathbb{A} = (A_1, \ldots, A_w), w \in \mathbb{N}$ of which one may be chosen as active controller in a particular time step. It should be noted that since $\mathbb{M}$ maps from $X = X^E \cap X^M$ back to X, it is actually a combined representation of world and agent, and in particular also value, dynamics. Such a model will almost never turn out to end as a perfect representation of the environment but in the most simple simulation experiments. Therefore, when handling predictions, evaluation of future possibilities or constructing anticipative action strategies, uncertainty is not an additional complexity but an inherent feature of the problem. On one hand, for most situations, very coarse and approximate action plans may turn out robust enough to suffice and on the other hand model-based forward planning control actions for a complex nonlinear environment is hard and solving this problem in every time step can be computationally very expensive, if not infeasible. Furthermore, a self-directed system may find many different more important tasks for its computational power, if can be confident about the performance of its learned reactive strategies which it can perform with significantly lower effort. Using its world model and past experience to model an approximation to Bellman's value function V^*, or local approximations thereof, taking into account its own as well as environmental states to calculate value by repeated value iteration over experiences or simulated episodes, in which the agent may or may not be in very different motivational states. This may again "color" the locally approximated value of different environmental and internal states in quite a different way. The benefits of such dual representations in context of predictive models is discussed in more detail in [9] and other works by the same author.

Following [8], if the approximated value function is used in conjunction with the world model to train the controllers A_i, which may form a hierarchical structure, the system essentially creates its own habitual response models respecting its subjective knowledge (including errors or overgeneralizations) of the

environment and itself, as well as experience of value which has been aggregated over possibly very different situations. On the other hand, carefully planned and evaluated online performance can be memorized for subsequent use as training data for more specialized controller models. In biological brains, very specific representations of different types of value, e.g. wanting and liking, are found [4]. A dual representation of aggregated and current cognitive evaluation of subjective value can also be a possibility for developing subjective self-models of the agent by comparing its experiences, explicit belief states and value approximations and evaluating its own memorized or habitualized actions.

III. Adaptive Artificial Motivational Systems in Optimization-based Agents

The first research on non-stationary motivational systems has been [11], where an evolutionary approach is taken to invent motivations, i.e. internal reward signal generators, that increase the overall performance of a learning agent w.r.t. basic reward intake. In particular, it could be shown that this approach can generate collections of subgoals for a given, predefined goal represented in terms of sparse reward.

In the following section, we discuss explicitely adaptive value systems in our framework and relate them to existing implementations of systems who incorporate complex representations of value.

A. Rewarding Reward Functions

The presented framework in [11], [10] defines the notion of an optimal reward function as a reward function that maximizes expected return, i.e. internal combined with external reward, over the distribution of all possible environments (albeit all possible here probably means already constrained to certain classes of quite regular and nice environments). In simulation experiments, they were able to evolve intrinsic as well as secondary extrinsic reward functions for an agent situated in a small discrete gridworld-type scenario.

This notion is also quite practical, as it allows for straightforward implementation in an already existing reinforcement learning agent. This approach is particular interesting when already more than one basic reward function is considered, for which an optimal configuration of additional reward functions is sought. In this case, an optimal reward function is required to reflect the mutual structure of the tasks represented by all considered basic reward functions, which can have very interesting consequences: (1) explicit representation of complex multi-component motivations, probably incorporating intrinsic as well as extrinsic components, e.g. curiosity towards states related to stimuli relevant to particular extrinsic motivations such as curiosity towards different possible food sources. (2) generalization of such complex motivational states to useful higher-order concepts of motivations. For such generalizations, the dynamic drive controller concept introduced in Sect.II.B. can serve to adjust the system's actual motivational configuration adaptively to the current state of the world.

B. Self-Adaptive Value Systems: Bootstrapping New Criteria for Behavior Selection and Self-Modification From Past Experiences

In Sect.II.B., we have defined a demand by means of a transformation from the agent state space x^A to its motivational state space $x^{M(A)}$. This may be represented by a nonlinear function $f^d : \mathbb{R}^n \to \mathbb{R}^q$. Such a function may be represented using static or dynamic parametric function approximators, e.g. neural network models, and optimized akin to [10]. Within our formal framework, this implies a time-dependent structure of the collection of motivational states $x^{M(A)}$, which can be realized using a modular structure of components (function approximators), where components may be added and non-basic components may be deleted. In any case, evaluating different structures/collections, which can be done offline using environmental and self-models, may take place using only x^A and r. Using the framework from [10], search for good representations of value is framed as an optimization problem which may be solved also using genetic algorithms, or other global search strategies. In an architecure which does not utilize an explicit world model, the task is however still possible to achieve using backtracking-type methods, e.g. the success-story-algorithm presented in [22].

Moreover, the adaptive parameters of the agent value system, the drive controller C, as well as the weighting parameters α, β, may be adapted in a similar way. By plugging a different set of parameters into a self-model and doing internal simulations starting from a designated situation, the agent can simulate how it would have performed in a particular episode it remembers. In that way, it is possible for the agent to modify what has been compared to a "personality" in order to be more successful. Summarizingly, the internal structure of the motivational system of the agent is envisioned to change by internal "what if?"-simulations or cognitive evaluations of situations within a structure which is conceptually separated from action selection, akin to the "non-action centered subsystem" in CLARION [9]. Since changing the set of internal reinforcers, i.e. reward functions, as well as adjustments to drive controller dynamics change the reward gradient pattern observed by the system, the optimal value function of agent/environment interaction is thereby also altered. In context of a dual action representation consisting of (1) a goal-directed cognitive planner and (2) a habitualized automatic actor system (or ensemble thereof) this has interesting consequences from a developmental perspective. Since all different value functions depend on innate value dynamics and the system continually tries to increase reward intake driven by innate value, adjustment of drive controller and weighting parameters might follow changes in the configuration of internal reinforcers (and vice versa), which may serve as a computational setting for the interrelation between requirements, tasks and goals on one hand, and on the other hand what has been called personality in the last paragraph.

A crucial point is the representation of these demands within the architecture of the control system. Modern neuroevolution techniques focussed on efficient manipulations of modular structures, e.g. NEAT [23], particularly in combination with evolutionary input varable selection methods may be well suited for this problem. They have, however, the drawback of resulting in intractable black-box models of motivational states. Dynamic fuzzy logic systems may provide a possibly for a tractable representation of motivational state dynamics, but since these models suffer from serious problems through an inherent approximation capability vs. interpretability tradeoff, whether they may express such states without becoming intractable is an question open for future research.

[2], [19] emphasized the importance of innate value systems from which current experience can be evaluated and new criteria can be bootstrapped. These value systems depend in their structure, functioning and actual implementation on three main dimensions of innate structure: morphology, internal dynamics and behavior [1]. From a designers point-of-view, it is therefore crucial for a complex self-development to be possible that a sufficient amount of innate knowledge w.r.t. morphology, internal dynamics and possible behaviors may be represented within the system from start. On the other hand, it may seriously speed up this developmental processes when the right "hints" for more complex representations of value are already present from the beginning. Many researchers have proposed hand-designed motivational states which may prove suitable as innate representations of value. For example, [9] describes the motivational system of the cognitive architecture CLARION, which encompasses a wide variety of motivational states formulated as drives. The set of primary desires is quite similar to that proposed in [15], [6] and represents basic needs for energy, avoidance of physical danger, and sleep. Additionally, a considerably sized set of high-level primary drives is suggested based on psychological findings, and the concept of a high-level drive is further justified. This set encompasses, among others, affiliation/belongingness, recognition/achievement, curiosity, conservation, autonomy and fairness. It seems interesting to investigate, (1) whether more complex high-level drives than a notion of curiosity based on internal model learning progress, such as the motivational concepts mentioned before, may emerge in self-adaptive value systems and (2) what kinds of derived simple or composite motivational states can emerge from a particular set of initial, innate value representations.

IV. Conclusion

A. Discussion

In this theoretical paper, artificial motivation systems have been examined from the viewpoint of self-referential control systems, a notion coined by [2] as a conceptual framework for studying developing self-modifying intelligent systems. We have discussed the reinforcement-learning-based framework for artificial motivation presented in [8], contrasted it with related approaches and derived a more general framework which may capture a wider range of interactions between action selection, motivational dynamics and cognitive processes, based on a combination of direct and indirect reinforcement learning. In particular, it was shown how such a dual representation can foster situated self-referential control in complex environments and may provide a useful architectural basis for the study of artificial emotions.

B. Future Work

In the future, we want to examine practical implementations of the ideas discussed in this paper. In particular, we are interested in transparent motivational systems whose functioning may be supervized by human operators/supervizors of an artificially motivated system using dynamic fuzzy logic systems. Moreover, we want to extend this framework for artificial motivational system towards representing elicitation and cognitive consequents of incorporating constructive models of artificial emotions.

References

[1] O. Sporns and E. Körner, *Towards a Theory of Thinking: Building Blocks for a Conceptual Framework*. Springer, 2010, ch. Value and Self-Referential Control: Necessary Ingredients for the Autonomous Development of Flexible Intelligence, pp. 323–335.

[2] E. Körner and G. Matsumoto, "Cortical architecture and self-referential control for brain-like computation," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 21, no. 5, pp. 121–133, 2002.

[3] W. Schultz, "Behavioral theories and the neurophysiology of reward." *Annu. Rev Psychol.*, vol. 57, pp. 87–115, 2006.

[4] K. C. Berridge and T. E. Robinson, "Parsing reward," *Trends in neurosciences*, vol. 26, no. 9, pp. 507–513, 2003.

[5] A. Sloman, *The Philosophy of Artificial Intelligence*. Oxford University Press, 1990, ch. Motives,Mechanisms,Emotions, pp. 217–234.

[6] J. Bach, *Principles of Synthetic Intelligence: Psi - An Architecture of Motivated Cognition*. New York and Oxford: Oxford University Press., 2009.

[7] ——, *Artificial General Intelligence*. Springer Berlin Heidelberg., 2011, ch. A Motivational System for Cognitive AI, pp. 232–242.

[8] M. Keramati and B. S. Gutkin, "A reinforcement learning theory for homeostatic regulation," in *Advances in Neural Information Processing Systems*, 2011, pp. 82–90.

[9] R. Sun, "Motivational representations within a computational cognitive architecture," *Computational Cognition*, vol. 1, pp. 91–103, 2009.

[10] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg, "Intrinsically motivated reinforcement learning: An evolutionary perspective," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 2, pp. 70–82, 2010.

[11] S. Singh, R. L. Lewis, and A. G. Barto, "Where do rewards come from?" in *31st Annual Conference of the Cognitive Science Society, Amsterdam*, 2009, pp. 2601–2606.

[12] A. Newell and H. Simon, *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

[13] C. Hull, *Principles of Behavior*. New York: Appleton-Century-Crofts., 1943.

[14] S. Reiss, *Who am I? The 16 Basic Desires that Motivate Our Actions and Define Our Personalities*. Berkley Trade., 2002.

[15] D. Dörner and K. Hille, "Artificial souls: Motivated emotional robots." in *IEEE Conference Proceedings, International Conference on Systems Man, and Cybernetics; Intelligent Systems for the 21st Century. (Vancouver, Canada)*, 1995.

[16] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990-2010)." *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[17] N. Chentanez, A. Barto, and S. Singh, "Intrinsically motivated reinforcement learning," in *Advances in neural information processing systems 17.*, 2004, pp. 1281–1288.

[18] J. Schmidhuber, "A possibility for implementing curiosity and boredom in model-building neural controllers." in *Proc. of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats.*, J. Meyer and S. Wilson, Eds. MIT Press/Bradford Books, 1991, pp. 222–227.

[19] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *Evolutionary Computation, IEEE Transactions on*, vol. 11, no. 2, pp. 265–286, 2007.

[20] G. L. Clore and J. R. Huntsinger, "How emotions inform judgment and regulate thought," *Trends in cognitive sciences*, vol. 11, no. 9, pp. 393–399, 2007.

[21] J. Schmidhuber, "Reinforcement learning in markovian and non-markovian environments," 1991.

[22] J. Schmidhuber, J. Zhao, and M. Wiering, "Shifting inductive bias with success-story algorithm, adaptive levin search, and incremental self-improvement," *Machine Learning*, vol. 28, no. 1, pp. 105–130, 1997.

[23] K. O. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evolutionary computation*, vol. 10, no. 2, pp. 99–127, 2002.