Chapter 15
Intrinsic Motivation for Truly Autonomous Agents

Ron Sun

15.1 Introduction

In order to deal with complexity, uncertainty, and unpredictability, which are inevitable in many real-world tasks and environments, agents need to be intrinsically motivated. Intrinsically motivated agents are those that have human-like (or animal-like) internal motivational processes, with internally generated, self-determined needs and preferences, which may or may not be influenced externally. It is the ability and the inclination of an agent (e.g., a human or a robot) to act autonomously, at its own discretion [6, 48]. For true autonomy necessary for dealing with highly complex, uncertain, or unpredictable environments, intrinsic motivation would be a highly desirable, or even necessary, part of being autonomous agents functioning in such environments. In highly complex, uncertain, or unpredictable environments, specific motivations and preferences cannot be easily pre-specified for a system from the outside, and thus intrinsic motivation is important for the sake of autonomy and for coping with such environments [18, 59].

In past work on intelligent agents, including past work on learning, planning, and problem solving for such agents, the need for intrinsic motivation has been down-played (although not completely ignored; more on this later). Thus, by now, the shortcomings of existent autonomous agent models and systems are quite evident, for example, with regard to their acceptance and their deployment in complex, uncertain, or unpredictable environments. Clearly, we need to seriously rethink some of these old approaches based on old (and often outdated) assumptions and methodologies, and move forward to the development of new, different, and better approaches, models, and theories, especially those that involve human-like intrinsic motivation.

Having intrinsic motivation is also important to achieving trust of autonomous agents and systems (such as autonomous robots) by humans (and by other autonomous agents and systems). In fundamentally unpredictable environments, a key aspect that one can be certain of is stable internal needs and preferences–that is, intrinsic moti-

R. Sun (✉)
Cognitive Sciences Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
e-mail: dr.ron.sun@gmail.com

H. A. Abbass et al. (eds.), *Foundations of Trusted Autonomy*, Studies in Systems, Decision and Control 117, https://doi.org/10.1007/978-3-319-64816-3_15

vation. Thus, in order to have trust and confidence in someone else, one has to have an understanding of what motivates the other [47, 48].

We may term human-like intrinsic motivation and autonomous choice of action (in accordance with intrinsic motivation) "free will". Self-determined intrinsic motivation (or "free will") in humans includes not only power, achievement, and other individualistic tendencies, but also adherence to social norms, affiliation with other individuals, and other tendencies related to social interactions and interdependencies ([48]; see details later). These motives are the results of evolution over a long period of human prehistory in the context of the struggles to survive within social groups. Real trust is trust among such "free willed" individuals. Limited, simpler forms of "trust" that one typically places on currently available machines such as self-driving automobiles or robotic vacuum cleaners (as they stand currently) cannot be construed as real trust (or full trust) and, I believe, is far from sufficient for the future. See, for example, Lee and See [23] or Abbass et al. [1] for characterization of such limited forms of trust. The question is: How do we move beyond that?

To achieve real trust, I believe that we need to delve into natural human tendencies to trust other individuals with intrinsic motivations that are similar to ours and similarly "free willed". Humans do have such tendencies, necessitated by their collective need for survival, evolved during their collective struggles to survive for tens of thousands of years. Such trust may start from predictability of behavior, as a result of similarly endowed (innate or acquired) motives. Understanding others' motivation leads to predictability of their behavior, which in turn leads to more complex and deeper forms of trust (e.g., involving affective or emotional processes). Only in this way, through understanding and exploiting such natural human tendencies, may we achieve truly autonomous agents, robots, and machines that may be given our real and full trust and that may also achieve real mutual trust amongst themselves.

Taking all of these issues into consideration, it is evident that we need to develop a deeper perspective on future autonomous systems, which should include intrinsic motivation in particular.

In the remainder of this chapter, first, background of some past work on human motivation is reviewed, as well as past work on computational cognitive architectures in relation to motivation. Then, a particular cognitive architecture (namely, the Clarion cognitive architecture) that is integrative and comprehensive and includes a more complete motivational subsystem is detailed, especially the interaction between its motivation and cognition [50]. Some examples of simulations using this cognitive architecture are described, which show briefly how this cognitive architecture integrates cognition and motivation and enables agents to function autonomously and appropriately. Some concluding remarks end this chapter.

15.2 Background

15.2.1 Previous Work on Intrinsic Human Motivation

It has been argued that, currently, many kinds of intelligent artifacts–autonomous agents, systems, and robots–are not truly autonomous, capable of dealing with complex, uncertain, and unpredictable environments independently, that is, truly autonomously. For one thing, they do not seem to possess independent, intrinsic motivations, needs, and preferences by themselves [48]. Notably, in highly complex, uncertain, or unpredictable environments, specific motivations and preferences cannot be easily pre-programmed; intrinsic motivation is therefore important to achieving autonomy and consequently crucial to successful coping in such environments [18, 59]. We need to make an effort to redress this current state of affairs.

Furthermore, we also need to address social interaction of and with autonomous agents, systems, and robots. For example, in social transactions, how and when one can place trust in such a system is a major issue, as mentioned earlier [24, 36, 39]. For humans, to truly trust and have confidence in someone else, one has to have an understanding of what motivates the other [47]. Having stable intrinsic motivation, as humans usually do, helps in this regard. For another example, social impasses may often result from incompatible motivations of multiple people (or agents); understanding each other's motivations may go a long way in helping to resolve such impasses [47].

Work on intrinsic human motivations has had a long history. Some particularly relevant work will be briefly discussed here [29, 34], in relation to our own theory of human motivation as embodied in the Clarion cognitive architecture mentioned earlier [46, 48, 50]. Understanding and replicating the human motivational subsystem can be highly beneficial to building autonomous intelligent agents and systems, because of its power, flexibility, and adaptability [48].

First of all, very early on, Murray [29] proposed a pertinent set of basic needs (i.e., primary drives in our terminology, as used in Clarion). Murray's proposal [29] included the need for conservance, the need for order, the need for retention, the need for acquisition, the need for inviolacy, and so on (note that these needs are included as or covered by primary drives in Clarion, as will be detailed later). Some other needs identified by Murray, such as contrarience, aggression, abasement, rejection, succorance, exposition, construction, and play, may not be fundamental needs (or primary drives) in our view—they are likely the results of more fundamental needs (i.e., primary drives) or their combinations. Murray's proposal also included some low-level (physiological, or viscerogenic in Murray's term) needs (which may be attributed to some combinations of low-level primary drives in Clarion).

More recently, Reiss [34] proposed another set of basic needs (i.e., primary drives), which was highly similar to Murray's, but with some differences. For example, as proposed by Reiss [34], there are the need for saving, the need for order, the need for family, the need for vengeance, the need for "idealism", the need for status, the need for acceptance, as well as the need for eating, the need for tranquility, the need for

physical exercises, the need for romance, and so on. (Again, these needs are included as or covered by primary drives in Clarion as will be detailed later.)

As alluded to above, in Sun [48, 50], a detailed model of human motivation (as embodied in Clarion) was presented. The Clarion cognitive architecture incorporates multiple, interacting subsystems. In particular, within the motivational subsystem, there are implicit drives and explicit goals (with goals being primarily determined based on drives). While some drives, denoting essential needs and desires, are primary and built-in, some other drives may be acquired and secondary. The primary drives include: Affiliation & Belongingness, Dominance & Power, Recognition & Achievement, Autonomy, Deference, Similance, Fairness, Honor, Nurturance, Conservation, Curiosity, as well as some low-level primary drives [48]. With its built-in mechanisms and processes, especially the motivational mechanisms, Clarion is able to capture, account for, and explain many psychological data and phenomena related to human motivation. There have been various efforts at verifying those drives through experiments and data analysis [34, 48]. See further discussions of Clarion below, and also see Sun [46, 50].

Relatedly, Schwartz's [40] 10 universal values, although addressing a different aspect of human behavior (i.e., human "values"), bear some resemblance to the essential needs (i.e., primary drives) identified above [48]. Moreover, each of these values can be derived from some primary drive or some combination of these primary drives [48].

McDougall [27] proposed a framework that was concerned with "instincts". Instincts, in our framework, refer to (more or less) evolutionarily hard-wired (i.e., innate) behavior patterns or routines that can be relatively easily triggered by pertinent stimuli in pertinent situations. As discussed earlier, basic needs (or primary drives as termed in Clarion) are essential driving forces of behaviors. Instincts are different from basic needs, because one does not have to follow instincts when there is no pertinent stimulus, and even when pertinent stimuli are present, one may be able to refrain from following instincts (at least more easily than from basic needs or primary drives). In other words, they are pre-set routines: while they are relatively easily triggered, they are not inevitable. McDougall listed the following instincts: imitation, emulation or rivalry, pugnacity/anger/resentment, sympathy, hunting, fear, appropriation/acquisitiveness, constructiveness, play, curiosity, sociability and shyness, secretiveness, cleanliness, modesty and shame, love, jealousy, parental love, …, and so on (see also [19]). As evident from the list above, many of these instincts are results of primary drives or basic needs (such as "curiosity" and "parental love"), or are derived, by some means, from primary drives or basic needs (such as "play" and "constructiveness"). Some other instincts are not because they do not represent basic needs (e.g., "hunting" or "jealousy"). (See more discussions of primary drives within Clarion later.)

There have also been some less psychologically validated models of motivation. Such models include Doerner's model and Sloman's model. In Sloman's motivational model [67], goals ("motives") are generated from a suite of modules ("generactivators"), each of which expresses a single "concern" (such as caring for dependents or removing damaged dependents). Each of these modules may search through a data-

base of beliefs; if it finds a match, a declarative representation of a goal (a "motive") is generated. On that basis, the resource management system takes goal representations and generates intentions for action. Although the model bears some resemblance to Clarion, the model has not been used to capture or explain psychological data in any detail. In addition, computationally speaking, searching through databases is cumbersome and may not be cognitively realistic.

Doerner [15] (see also Bach [5]) described the PSI theory, which included internal deficits, displeasure signals (due to deficits), negative reinforcement (from displeasure signals), urges, goals, action learning through random exploration (based on reinforcement), and so on. At an abstract level, the model is similar to Clarion to some extent [48, 50], but it appears less psychologically grounded or validated. In addition, its computational mechanisms appear less well developed algorithmically.

15.2.2 Previous Work on Cognitive Architectures

It has been suggested [30] that cognitive theories (including computational cognitive models) should be developed that satisfy multiple criteria, in order to avoid theoretical myopia. There have been steady developments of generic computational cognitive models, that is, cognitive architectures, for the past three decades since that seminal suggestion.

Early cognitive architectures often took the form of production systems and were (more or less) concerned with various psychological phenomena [20]. However, other forms of cognitive architectures have also been developed over the years — they may be in the form of a connectionist model, a constraint satisfaction network, a hybrid system of different models, and so on. Some of them may be more concerned with applications to building artificial systems than capturing and explaining empirical psychological phenomena.

Computational cognitive architectures provide the best hope for integrated systems that incorporate not just cognitive capabilities, but also motivation, emotion, personality, and many other capacities and capabilities needed for an autonomous agent. In particular, computational cognitive architectures based firmly on psychological data and findings and thus well-grounded empirically can be especially illuminating– they provide a glimpse into how human minds work, for example, in terms of the interaction between cognition and motivation, as well as their interaction with the environments (simple or complex). The human mind provides the best example of a truly autonomous intelligent system, and thus can lead to better understanding of intelligence and autonomy. Such cognitive architectures, like humans on which they are based, are capable of being truly autonomous, because they include a wide range of cognitive, motivational, and other capabilities and these capabilities function together to cope with different tasks and environments. Let us look into three examples, in chronological order.

Soar, the first proposed cognitive architecture, has been developed over the past thirty years, based essentially on a production system model. It has mostly been used for the purpose of building application systems [21, 30, 35]. In Soar, based on the framework of a state space and operators for searching the state space, decisions are made by different productions proposing different operators, when there is a goal on a goal stack. When a sequence of productions leads to achieving a goal, chunking occurs, which creates a single production that summarizes the process (using explanation-based learning). However, it lacks sophisticated motivational structures and processes. In addition, a large amount of initial (a priori) knowledge about states and operators is required for Soar to work.

Another series of cognitive architectures were also proposed fairly early on: in particular, ACT* and ACT-R [3]. ACT* is made up of declarative knowledge (captured in a semantic network) and procedural knowledge (captured in a production system). Procedural knowledge (in productions) is acquired through "proceduralization" of declarative knowledge, modified through use by generalization and discrimination (i.e., specialization), and have strengths associated with them (which are used for firing). ACT-R is a descendant of ACT*, in which procedural learning is limited to production formation through mimicking and production firing is based on log odds of success. There have been some later additions to ACT-R, including visual and motor modules, but there have not been any sufficiently complex motivational structures.

Clarion has been a comprehensive cognitive architecture [45, 46, 50]. The Clarion cognitive architecture, as mentioned earlier, consists of multiple, interacting subsystems. It is also distinguished from other existing cognitive architectures by its focus on the separation and the interaction of implicit and explicit knowledge and processes (in these different subsystems, respectively). More importantly, in relation to motivational issues, compared with other cognitive architectures, Clarion is distinguished by the fact that it contains built-in motivational constructs and built-in metacognitive constructs. These features are not commonly found in other existing cognitive architectures. Nevertheless, these features are crucial to the cognitive architecture, as they capture important elements in the interaction between an agent and its physical and social world [50]. With these mechanisms, especially the motivational and metacognitive mechanisms, Clarion attempts to explain their functioning in concrete computational terms.

15.3 A Cognitive Architecture with Intrinsic Motivation

15.3.1 Overview of Clarion

Clarion provides structural and algorithmic specifications of a wide range of generic psychological processes. In particular, Clarion accounts for basic human motives, which provide the underlying basis for behavior. This emphasis on human moti-

vation facilitates the integration of general cognitive capacities with considerations of motivation (as well as personality, emotion, culture, sociality, and so on) in a comprehensive and unified theory/model.

Only a sketch of Clarion can be presented below; the vast majority of technical details are omitted due to the page limit. See Fig. 15.1 for the overall structure of Clarion.

As shown by the figure, Clarion consists of a number of subsystems: the action-centered subsystem (denoted as the ACS), the non-action-centered subsystem (denoted as the NACS), the motivational subsystem (the MS), and the metacognitive subsystem (the MCS). The role of the action-centered subsystem is to control actions (regardless of whether they are for external physical movements or for internal mental operations), utilizing and maintaining procedural knowledge. The role of the non-action-centered subsystem is to maintain and utilize declarative knowledge. The role of the motivational subsystem is to provide underlying motivations for perception, action, and cognition (in terms of providing impetus and feedback). The role of the metacognitive subsystem is to monitor, direct, and modify the operations of the other subsystems dynamically.

Each of these interacting subsystems consists of two "levels" of representations (i.e., a dual-representational structure, as theoretically posited in [45]). Generally speaking, in each subsystem, the "top level" encodes explicit knowledge[1]

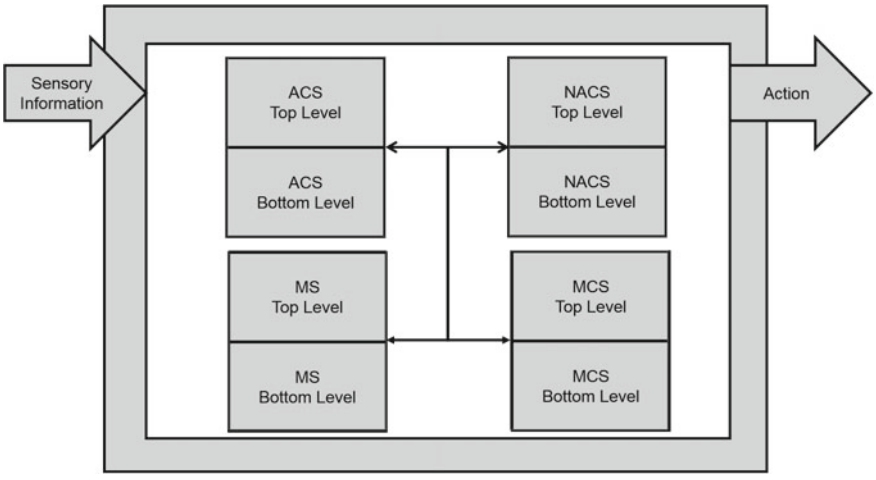

Fig. 15.1 The Clarion cognitive architecture. The subsystems of Clarion are shown. The major information flows are shown with arrows. ACS stands for the action-centered subsystem. NACS stands for the non-action-centered subsystem. MS stands for the motivational subsystem. MCS stands for the metacognitive subsystem

[1]Roughly speaking, explicit knowledge is directly consciously accessible (i.e., conscious or potentially conscious), while implicit knowledge is consciously inaccessible directly. Explicit processes involve explicit knowledge, while implicit processes involve implicit knowledge. The distinction has been based on voluminous empirical findings in many domains, but involves some nuances and some controversies. See [45, 50] for details.

(using symbolic/localist representations) and the "bottom level" encodes implicit knowledge (using distributed representations [32, 37]). The two levels interact, for example, by cooperating in action decision making, through integration of the action recommendations from the two levels of the ACS respectively, as well as by cooperating in learning through a "bottom-up" and a "top-down" learning process [45, 55].

Existing theories tend to confuse implicit and explicit processes; hence the "perplexing complexity" [43]. In contrast, Clarion generally separates and integrates implicit and explicit processes in each of its subsystems. With such a framework, Clarion can provide better explanations of empirical findings in a wide range of domains (for details, see [17, 45, 55]).

15.3.2 The Action-Centered Subsystem

The ACS captures the process of human action decision making as follows: Observing the current (observable) state of the world (including one's own motivational state), the two levels within the ACS (implicit or explicit) make their separate action decisions in accordance with their respective procedural knowledge (implicit or explicit), and their outcomes are "integrated". Thus, a final selection of an action is made and the action is then performed. The action changes the world in some way. Comparing the changed state of the world with the previous state, the person learns. The cycle then repeats itself.

In this subsystem, the bottom level consists of "action neural networks" encoding implicit knowledge (involving distributed representations [37]), and the top level consists of "action rules" encoding explicit knowledge (using symbolic/localist representations).

At the bottom level of the ACS, using an action neural network, actions are selected based on their Q values. At each step, given state x, the Q values of all the actions in that state (i.e., $Q(x, a)$ for all a's) are computed in parallel. Then the Q values are used to decide stochastically on an action to be performed, through a Boltzmann distribution of Q values:

$$p(a|x) = \frac{e^{\frac{Q(x,a)}{\tau}}}{\sum_i e^{\frac{Q(x,a_i)}{\tau}}}$$

where $p(a|x)$ is the probability of selecting action a, τ (temperature) controls the degree of randomness of action decision making, and i ranges over all possible actions. (This is known as Luce's choice axiom [61].)

For capturing learning of implicit knowledge at the bottom level (i.e., the Q values), the Q-learning algorithm [61], a reinforcement learning algorithm, may be applied. With this algorithm, Q values are gradually tuned through successive

updating of a neural network, which enables reactive sequential behavior to emerge through trial-and-error interaction with the world (for details, see [45, 61]).

For capturing learning of explicit knowledge at the top level (i.e., action rules), a variety of algorithms may be applied, including the Rule-Extraction-Refinement (RER) algorithm [11] for a "bottom-up" learning process that relies on implicit knowledge from the bottom level to learn explicit knowledge at the top level [45]. In the reverse direction, "top-down" learning can also occur.

For stochastic selection of the outcomes of the two levels, at each step, each level (or a component within) is selected with a certain probability. There exists some psychological evidence for such intermittent use of rules [45]. The selection probabilities may be variable, determined by the metacognitive subsystem (by its processing mode module; more later; [50]).

15.3.3 The Non-Action-Centered Subsystem

The NACS is for dealing with declarative knowledge (which is not action-centered). It stores such knowledge in a dual representational form (the same as in the ACS): that is, in the form of explicit "associative rules" (at the top level), and in the form of implicit "associative memory networks" (at the bottom level). Its operation is under the control of the ACS and in the service of the ACS.

First, at the bottom level of the NACS, associative memory networks encode implicit declarative knowledge. Associations are formed by mapping an input pattern to an output pattern (e.g., using Backpropagation networks or Hopfield networks [37]).

Second, at the top level of the NACS, explicit declarative knowledge is stored. As in the ACS, each "chunk" node (denoting a concept) at the top level is linked to its corresponding microfeature nodes present at the bottom level. Additionally, in the top level, links between chunk nodes encode explicit associative rules. Explicit associative rules may be learned in a variety of ways [50].

As in the ACS, top-down or bottom-up learning may take place in the NACS, either to extract explicit knowledge at the top level from the implicit knowledge at the bottom level, or to assimilate the explicit knowledge of the top level into the implicit knowledge at the bottom level.

With the interaction of the two levels, the NACS carries out rule-based, similarity-based, and constraint-satisfaction-based reasoning (details can be found in [17, 50]). Their interaction enables the NACS to capture much of human reasoning [50].

15.3.4 The Motivational Subsystem

The MS is a critical part of the cognitive architecture. It is concerned with why an individual does what he/she does. The importance of the MS to the ACS lies in

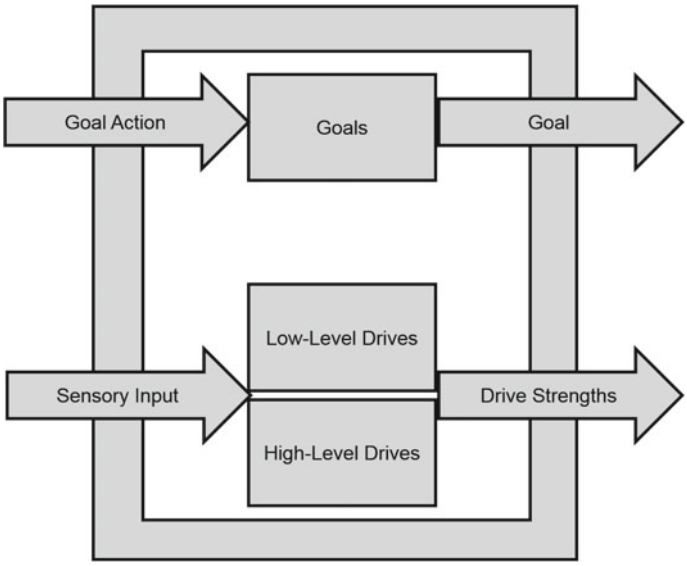

Fig. 15.2 The basic structure of the motivational subsystem

the fact that it provides the context in which goals and reinforcements of the ACS are determined. It thereby influences the working of the ACS (and by extension, the working of the NACS).

A dual motivational representation is in place in the MS. The explicit goals at the top level of the MS (such as "find food"), which are essential to the working of the ACS, may be generated based on implicit drives at the bottom level of the MS (e.g., "hunger"). See Fig. 15.2. For justifications, see [48].

At the bottom level of the MS, primary drives are those motives essential to an individual and most likely built-in (hard-wired) to a significant extent to begin with (i.e., they are "intrinsic"). Low-level primary drives (concerning mostly physiological needs) include: food, water, reproduction, and so on. Beyond low-level primary drives, there are also high-level primary drives: for example, achievement and recognition, affiliation and belongingness, dominance and power, fairness, autonomy, and so on (see [29, 34, 48, 58, 62]).[2] These primary drives have been justified in prior writings (as cited above).[3] See Table 15.1 for their specifications. On the basis of primary drives, secondary (derived) drives may be acquired.

[2]Note that a generalized notion of "drive" is adopted in Clarion. As discussed in [48], it is a generalized notion that transcends controversies surrounding the stricter notions of drive [18].

[3]Briefly, this set of hypothesized primary drives bears close relationships to Murray's needs [29], Reiss's motives [34], Schwartz's universal values [40], and so on. The prior justifications of these frameworks may be applied, to a significant extent, to this set of drives as well (see [25, 29, 34, 48]).

Table 15.1 Descriptions of the primary drives

Drives	Specifications
Food	The drive to consume nourishment
Water	The drive to consume liquid
Sleep	The drive to rest
Reproduction	The drive to mate
Avoiding Danger	The drive to avoid situations that have the potential to be harmful
Avoiding Unpleasant Stimuli	The drive to avoid situations that are physically (or emotionally) uncomfortable or negative in nature
Affiliation & belongingness	The drive to associate with other individuals and to be part of social groups
Dominance & power	The drive to have power over other individuals
Recognition & achievement	The drive to excel and be viewed as competent
Autonomy	The drive to resist control or influence by others
Deference	The drive to willingly follow or serve a person of a higher status
Similance	The drive to identify with other individuals, to imitate others, and to go along with their actions
Fairness	The drive to ensure that one treats others fairly and is treated fairly by others
Honor	The drive to follow social norms and codes and to avoid blames
Nurturance	The drive to care for, or attend to the needs of, others who are in need
Conservation	The drive to conserve, to preserve, to organize, or to structure (e.g., one's environment)
Curiosity	The drive to explore, to discover, and to gain new knowledge

Table 15.2 Approach versus avoidance primary drives

Approach drives	Avoidance drives	Both
Food	Sleep	Affiliation & belongingness
Water	Avoiding danger	Similance
Reproduction	Avoiding Unpleasant Stimuli	Deference
Nurturance	Honor	Autonomy
Curiosity	Conservation	Fairness
Dominance & Power		
Recognition & Achievement		

Some of these primary drives are approach-oriented, while others are avoidance-oriented. This distinction has been argued by many (e.g., [12, 16, 43]). The approach system is sensitive to cues signaling rewards, and results in active approach. The avoidance system is sensitive to cues of punishment, and results in avoidance, characterized by anxiety or fear. See Table 15.2 for this division of drives.

The processing of these drives within the bottom level of the MS involves a number of modules [50]. In particular, the core drive module determines drive strengths (using

neural networks) based roughly on:

$$ds_d = gain_d \times stimulus_d \times deficit_d + baseline_d$$

where ds_d is the strength of drive d, $gain_d$ is the gain for drive d, $stimulus_d$ is a value representing how pertinent the current situation is to drive d, $deficit_d$ indicates the perceived deficit in relation to drive d (which represents an individual's internal inclination toward activating drive d), and $baseline_d$ is the baseline strength of drive d. The justifications for this may be found in the literature [50, 58, 60].

Motivational adaptation (learning) is also possible and has been tackled [50]. In addition, new drives (termed "derived drives") may be acquired. They may be gradually acquired through some kind of "conditioning", or may be externally set through externally provided instructions, on the basis of primary drives.

15.3.5 The Metacognitive Subsystem

Metacognition refers to active monitoring and consequent regulation and orchestration of one's own psychological processes [26, 33]. In Clarion, the MCS is closely tied to the MS. The MCS monitors, controls, and regulates other processes [42]. Control and regulation may be in the forms of setting goals (which are then used by the ACS) on the basis of drives, generating reinforcement signals for the ACS for learning (on the basis of drives and goals), interrupting and changing ongoing processes in the ACS and the NACS, setting essential parameters of the ACS and the NACS, and so on.

Structurally, this MCS may be divided into a number of functional modules, including:

- the goal module,
- the reinforcement module,
- the processing mode module.
- the input filtering module,
- the output filtering module,
- the parameter setting module (for setting learning rates, temperatures, etc.),

and so on. See Fig. 15.3.

For instance, the goal module selects goals to pursue (by the ACS). In order to select a new goal, it first determines goal strengths, based on information from the MS (e.g., the drive strengths) and the current sensory input. Then, a new goal is stochastically selected on the basis of the goal strengths (e.g., using a Boltzmann distribution). For arguments in support of goal setting on the basis of implicit motives (i.e., drives), see, for example, Tolman [59] and Deci [13]. In the simplest case, the following calculation is performed:

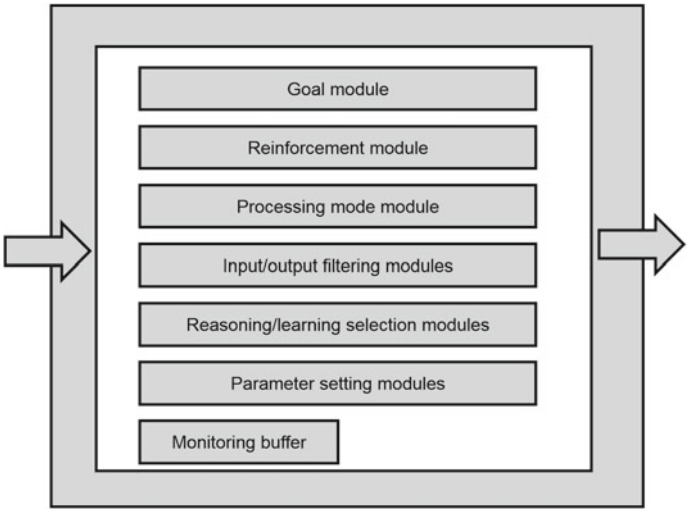

Fig. 15.3 The main modules within the metacognitive subsystem

$$gs_g = \sum_d relevance_{d,s\to g} \times ds_d$$

where gs_g is the strength of goal g, $relevance_{d,s\to g}$ is a measure of how relevant drive d is to goal g with regard to current situation s (which represents the support that drive d provides to goal g), and ds_d is the strength of drive d (from the MS). Once calculated, the goal strengths are turned into a Boltzmann distribution (as discussed earlier) and the new goal is chosen stochastically from that distribution.

For another instance, the processing mode module determines the probability of each component (a level or a component within) for the integration of outcomes from the two levels of the ACS (see the discussion of the ACS earlier). These probabilities may be determined through the notion of "probability matching": the probability of selecting a component is determined based on the relative success ratio of that component (see [45, 50] for details). However, these probabilities may be modulated multiplicatively by another parameter: the strength of avoidance-oriented primary drives (which corresponds to "anxiety" [65, 66]; see more details below).

15.4 Some Examples of Simulations

Clarion, as a framework, has been justified and validated extensively on the basis of psychological data and their simulations; see, for example, Sun [45, 50] for summaries of such justifications and validations. In particular, Clarion has been successful in simulating, accounting for, and explaining a wide variety of psychological

data. For example, a number of well-known skill learning tasks have been captured, simulated, and explained using Clarion that span the spectrum ranging from simple reactive skills to complex cognitive skills. Simulations have also been done with reasoning tasks, metacognitive tasks, and motivational tasks, as well as social interaction tasks, all of which are important to autonomous intelligent agents. While accounting for various psychological data, Clarion provides explanations that shed new light on relevant phenomena, especially on the basis of motivation.

Let us look into an example of social simulation involving agents with intrinsic motivation (as originally described in [51]). A significant shortcoming of many computational social simulations is that they assume very rudimentary cognition/psychology on the part of agents. Although agents are often characterized as being "cognitive", there has been relatively scarce effort that carefully captures human psychology in detailed, process-based, and quantitative ways. Models of agents have frequently been custom-tailored to the task at hand, often amounting to a set of highly domain-specific rules. Although such an approach may be adequate for achieving some limited objectives of some specific simulations, it falls short intellectually [53, 54]: It not only limits the realism of social simulations, but also precludes the possibility of fully tackling the question of the micro-macro link [2, 38, 44] in terms of psychological-social interaction, for example, how the social emerges from the psychological [44, 47, 51].

Thus, let us first look into an existing social simulation as an illustration. In the work of Cecconi and Parisi [10], social groups (tribes) were simulated. In these groups, to survive and to reproduce, an agent must possess resources. A group in which each agent uses only its own resources is said to adopt an individual survival strategy. However, in some other groups, resources may be transferred among agents—such a group is said to adopt a social survival strategy. For instance, the "central store" (CS) is a mechanism to which all the individuals in a group transfer (part of) their resources. The resources collected by the CS can be redistributed to the members of the group in some fashion [10].

In Cecconi and Parisi [10], a number of simulations were conducted comparing groups adopting individual strategies with groups adopting CS strategies. Agents survived and reproduced differentially based on the quantity of food that they were able to consume. This task has the potential for exploring a wide range of issues, ranging from individual behaviors to social institutions, and from individual learning to evolution.

However, in that work, there was very little in the way of psychological processes. Such work needs a better understanding, and better models, of the individual mind, because only on the basis of that understanding, better understanding of aggregate processes can be developed. Accurate, detailed cognitive/psychological models may provide better grounding for understanding multi-agent social phenomena, by incorporating realistic constraints, capabilities, and tendencies of individual agents. This point was argued in Sun [44]. In Axelrod [4], it was shown that even adding a cognitive factor as simple as memory of past several events into an agent model can completely alter the dynamics of social interaction (e.g., in the iterated prisoner's dilemma).

Thus, we conducted simulations based on the Clarion cognitive architecture. In our simulation, the world was made up of a 200×200 grid. Each of these 40,000 locations might contain (at most) one food item. At the beginning and every 40 cycles, the grid was replenished: Randomly selected locations were restocked with food items (until the grid had 2400 food items). A more benign condition, in which 3600 locations contained one food item each, and a harsher condition, in which 1200 locations contained one food item each, were also tested. A food item contained 50 energy units.

Each agent began with 60 units of energy, and consumed one unit of energy per cycle. Each agent lived for a maximum of 350 cycles, but might die early due to lack of food. There were initially 120 agents to begin with, and the number of agents fluctuated due to birth and death, within the bound of a maximum of 120 agents.

At each moment, each agent was in a particular location on the grid. It faced a certain direction (north, south, east, or west). Each received inputs regarding the location of the nearest food. Each agent could generate an action output: either (1) turn 90 degrees right, (2) turn 90 degrees left, (3) move forward, (4) pick up food and contribute a portion, (5) pick up food and keep all of it, or (6) reproduce.

As in the previous simulations, procreation was asexual; procreation occurred if an agent had reached 120 energy units or more, and there were fewer than the maximum number of agents in the world. The parent handed out 60 energy units to the child upon its birth. The child inherited its parent's internal makeup, although when a child was spawned, there was a 10% chance of minor mutation.

When a central store was involved, an agent was required to contribute 20 energy units to the central store when it picked up a food item (50 energy units). At each cycle, agents with 10 or less energy might receive 5 energy units each from the central store; up to a maximum of 10% of the agent population might get energy from the central store at each cycle. Each agent, when picking up a piece of food, decided whether to contribute to the central store or not. There were three variations on cheater detection and punishment, ranging from full detection and full punishment to no detection and no punishment.

Each agent had three intrinsic drives: food, reproduction, and honor. They competed to influence behavior (action) through determining the current goal (e.g., to pursue food or reproduction, or to contribute or not to the CS). The internal reinforcements for their actions were determined based on their drives and goals, as well as the state of the world.

The results of the simulation demonstrated effects of motivational factors (which were not investigated in the previous simulations). As predicted, motivational factors had a significant effect on the outcome of the simulation. In this regard, "gains" was a variable created for the sole purpose of analysis; it consolidated the three drive gain parameters (for food, reproduction, and honor, respectively) into one. There were eight values, ranging from "All 0.5" to "All 1.0"; for example, "Honor 0.5" meant that the gain parameter of the Honor drive was 0.5 and all the other drive gains were 1.0 (see Fig. 15.4 for the complete list).

Examining the results in Fig. 15.4, there was a significant effect of "gains" on lifespan. Generally speaking, more emphasis on food (a higher drive gain for food) led

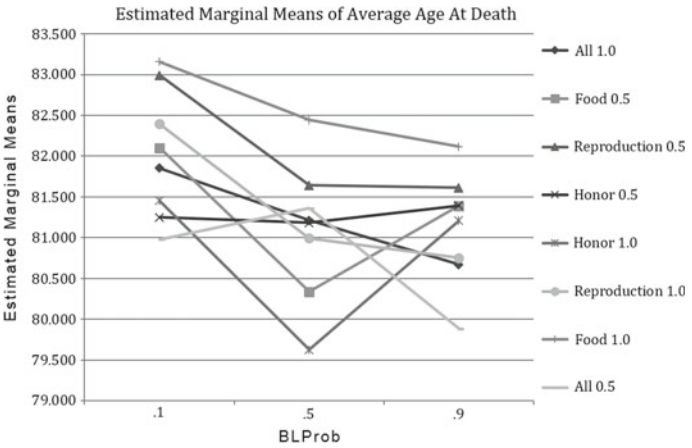

Fig. 15.4 The effect of gains on lifespan. The y-axis represents lifespan. The x-axis represents probability of using the bottom level. The different lines represent different drive gain settings

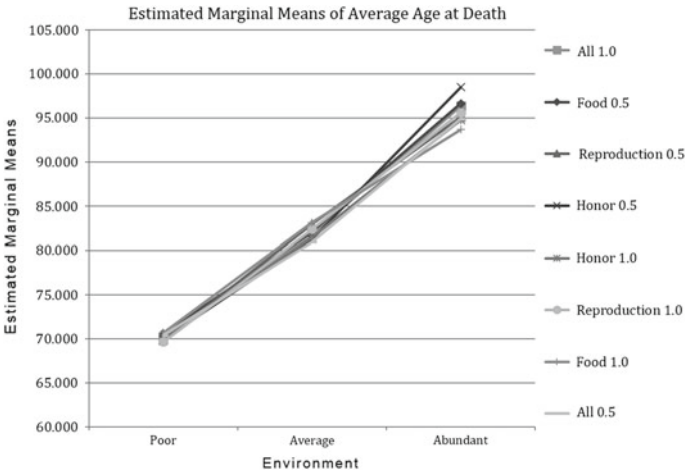

Fig. 15.5 The interaction of gains and environment on lifespan. The y-axis represents lifespan. The x-axis represents environment. The different lines represent different drive gain settings

to better performance (e.g., "Food 1.0" or "Reproduction 0.5"). Reduced emphasis on food generally led to worse performance (e.g. "Food 0.5" or "Honor 1.0").

There was also a significant interaction between "gains" and environment on lifespan, as shown in Fig. 15.5. An interpretation of this result was this: In a more benign environment, less focus on honor (e.g., "Honor 0.5") helped survival, but in a harsh environment, drive focuses did not make much difference because one had to focus only on food in order to survive.

Other related simulations of motivation or motivation-cognition interaction may be found in Wilson et al. [65, 66], especially in relation to effects of anxiety [7, 8, 22]. A model of personality was also developed on the basis of motivation within Clarion; see, for example, Sun and Wilson [56] (see also [14, 28, 31, 41]). Further, a model of moral judgment was developed on the basis of motivation; see [9, 49]. A model of emotion was also developed on the basis of motivation within Clarion; see Sun et al. [57] (see also [52, 63, 64]).

15.5 Concluding Remarks

The Clarion project addresses essential human-like motivational processes, mechanisms, structures, and representations necessary for a comprehensive cognitive architecture. The need for implicit drive representations, as well as explicit goal representations, has been hypothesized. Drive representations consist of primary drives (both low-level and high-level primary drives), as well as derived (secondary) drives. On the basis of drives, explicit goals may be generated on the fly during an agent's interaction with various situations, which in turn guide action selection.

The afore-discussed motivational representations and their resulting dynamics help to make a computational cognitive architecture more complete and functioning in a more psychologically realistic way. I believe that this constitutes a requisite step forward in making computational cognitive architectures more realistic models of the human mind taking into considerations all of its complexity and intricacy, especially in terms of its complex motivational dynamics. It is highly relevant to building truly autonomous and trust-worthy computational agents capable of functioning in complex, uncertain, and unpredictable environments. Note that what I emphasize here is human-like full autonomy and human-like trust.

Significant future challenges in furthering this line of work exist, including, for example, applying this framework to the building of intelligent application systems that can display intelligent behavior with more robustness, flexibility, and versatility. Another significant challenge is to further validate, through empirical work (especially psychological empirical work), this framework and its implications for understanding human motivation and trust (in addition to building intelligent agents). Many more experiments, simulations, and tests will be needed and shall be pursued in the future.

Acknowledgements This work has been supported in part by ONR grants N00014-08-1-0068 and N00014-13-1-0342.

References

1. H.A. Abbass, G. Leu, K. Merrick, A review of theoretical and practical challenges of trusted autonomy in big data. IEEE Access **4**, 2809–2830 (2016)
2. J. Alexander, B. Giesen, R. Munch, N. Smelser (eds.), *The Micro-Macro Link* (University of California Press, Berkeley, CA, 1987)
3. J.R. Anderson, C.J. Lebiere, *The Atomic Components of Thought* (Lawrence Erlbaum Associates, Mahwah, NJ, 1998)
4. R. Axelrod, *The Evolution of Cooperation* (Basic books, New York, 1984)
5. J. Bach, *Principles of Synthetic Intelligence* (Oxford University Press, New York, 2009)
6. G. Baldassarre, M. Mirolli (eds.), *Intrinsically Motivated Learning in Natural and Artificial Systems* (Springer, Berlin-Heidelbergr, 2013)
7. S.L. Beilock, T.H. Carr, On the fragility of skilled performance: What governs choking under pressure? J. Exp. Psychol. Gen. **130**(4), 701–725 (2001)
8. S.L. Beilock, C.A. Kulp, L.E. Holt, T.H. Carr, More on the fragility of performance: Choking under pressure in mathematical problem solving. J. Exp. Psychol. Gen. **133**(4), 584–600 (2004)
9. S. Bretz, R. Sun, Two models of moral judgement (2016). Submitted
10. F. Cecconi, D. Parisi, Individual versus social survival strategies. J. Artif. Soc. Soc. Simul. **1**(2), 1–17 (1998)
11. A. Clark, A. Karmiloff-Smith, The cognizer's innards: A psychological and philosophical perspective on the development of thought. Mind Lang. **8**(4), 487–519 (1993)
12. L.A. Clark, D. Watson, *Handbook of personality: Theory and research, chapter Temperament: A new paradigm for trait psychology* (Guilford Press, New York, 1999), pp. 399–423
13. E. Deci, *Personality: Basic Issues and Current Research, chapter Intrinsic motivation and personality* (Prentice Hall, Englewood Cliffs, NJ, 1980), pp. 35–80
14. J.M. Digman, Personality structure: Emergence of the five-factor model. Annu. Rev. Psychol. **41**(1), 417–440 (1990)
15. D. Dörner, The mathematics of emotions. *Proceedings of the Fifth International Conference on Cognitive Modeling* (2003), pp. 75–79
16. J.A. Gray, N. McNaughton, *The Neuropsychology of Anxiety: An Enquiry into the Function of the Septo-hippocampal System* (Oxford University Press, New York, 2000)
17. S. Hélie, R. Sun, Incubation, insight, and creative problem solving: A unified theory and a connectionist model. Psychol. Rev. **117**(3), 994–1024 (2010)
18. C.L. Hull, *Principles of Behavior: An Introduction to Behavior Theory* (D. Appleton-Century Company, 1943)
19. W. James, *The Principles of Psychology* (Dover, New York, 1890)
20. D. Klahr, P. Langley, R. Neches, *Production System Models of Learning and Development* (MIT press, Cambridge, MA, 1989)
21. J.E. Laird, *The Soar Cognitive Architecture* (MIT Press, Cambridge, MA, 2012)
22. A.J. Lambert, B. Keith Payne, L.L. Jacoby, L.M. Shaffer, A.L. Chasteen, S.R. Khan, Stereotypes as dominant responses: On the social facilitation of prejudice in anticipated public contexts. J. Pers. Soc. Psychol. **84**(2), 277–295 (2003)
23. J.D. Lee, K.A. See, Trust in automation: Designing for appropriate reliance. Hum. Factors J. Hum. Factors Ergon. Soc. **46**(1), 50–80 (2004)
24. R.J. Lewicki, D.J. McAllister, R.J. Bies, Trust and distrust: New relationships and realities. Acad. Manag. Rev. **23**(3), 438–458 (1998)
25. A.H. Maslow, A theory of human motivation. Psychol. Rev. **50**(4), 370–396 (1943)
26. G. Mazzoni, T.O. Nelson (eds.), *Metacognition and Cognitive Neuropsychology: Monitoring and Control Processes* (Erlbaum, Mahwah, NJ, 1998)
27. W. McDougall, *An Introduction to Social Psychology* (Methuen & Co., London, 1936)
28. W. Mischel, Y. Shoda, Reconciling processing dynamics and personality dispositions. Annu. Rev. Psychol. **49**, 229–258 (1998)
29. H.A. Murray, *Explorations in Personality* (Oxford University Press, New York, 1938)

30. A. Newell, *Unified Theories of Cognition* (Harvard University Press, Cambridge, MA, 1990)
31. S.J. Read, B.M. Monroe, A.L. Brownstein, Y. Yang, G. Chopra, L.C. Miller, Virtual personalities II: A neural network model of the structure and dynamics of human personality. Psychol. Rev. **117**(1), 61–92 (2010)
32. A.S. Reber, Implicit learning and tacit knowledge. J. Exp. Psychol. Gen. **118**(3), 219–235 (1989)
33. L.M. Reder, *Implicit Memory and Metacognition* (Erlbaum, Mahwah, NJ, 1996)
34. S. Reiss, Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. Rev. Gen. Psychol. **8**(3), 179–193 (2004)
35. P.S. Rosenbloom, J. Laird, A. Newell, *The SOAR Papers: Research on Integrated Intelligence* (MIT Press Cambridge, MA, 1993)
36. D.M. Rousseau, S.B. Sitkin, R.S. Burt, C. Camerer, Not so different after all: A cross-discipline view of trust. Acad. Manag. Rev. **23**(3), 393–404 (1998)
37. D.E. Rumelhart, J.L. McClelland, *Parallel Distributed Processing: Explorations in the Microstructures of Cognition* (MIT Press, Cambridge, MA, 1986)
38. R.K. Sawyer, Artificial societies multiagent systems and the micro-macro link in sociological theory. Sociol. Methods Res. **31**(3), 325–363 (2003)
39. K.E. Schaefer, D.R. Billings, J.L. Szalma, J.K. Adams, T.L. Sanders, J.Y. Chen, P.A. Hancock, *A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction* (Human Factors, to appear, Technical report, 2016)
40. S.H. Schwartz, Are there universal aspects in the structure and contents of human values? J. Soc. Issues **50**(4), 19–45 (1994)
41. Y. Shoda, W. Mischel, *Connectionist models of social reasoning and social behavior, chapter Personality as a stable cognitive-affective activation network: Characteristic patterns of behavior variation emerge from a stable personality structure* (Lawrence Erlbaum Associates Inc, Mahwah, NJ, 1998), pp. 175–208
42. H.A. Simon, Motivational and emotional controls of cognition. Psychol. Rev. **74**(1), 29–39 (1967)
43. L.D. Smillie, A.D. Pickering, C.J. Jackson, The new reinforcement sensitivity theory: Implications for personality measurement. Pers. Soc. Psychol. Rev. **10**(4), 320–335 (2006)
44. R. Sun, Cognitive science meets multi-agent systems: A prolegomenon. Philos. Psychol. **14**(1), 5–28 (2001)
45. R. Sun, *Duality of the Mind: A Bottom-up Approach Toward Cognition* (Lawrence Erlbaum Associates, Mahwah, NJ, 2002)
46. R. Sun, A tutorial on CLARION 5.0. Technical report, Cognitive Science Department, Rensselaer Polytechnic Institute, 2003
47. R. Sun (ed.), *Cognition and Multi-agent Interaction: From Cognitive Modeling to Social Simulation* (Cambridge University Press, 2006)
48. R. Sun, Motivational representations within a computational cognitive architecture. Cogn. Comput. **1**(1), 91–103 (2009)
49. R. Sun, Moral judgment, human motivation, and neural networks. Cogn. Comput. **5**(4), 566–579 (2013)
50. R. Sun, *Anatomy of the Mind: Exploring Psychological Mechanisms and Processes with the Clarion Cognitive Architecture* (Oxford University Press, 2016)
51. R. Sun, P. Fleischer, A cognitive social simulation of tribal survival strategies: The importance of cognitive and motivational factors. J. Cogn. Cult. **12**(3–4), 287–321 (2012)
52. R. Sun, R.C. Mathews, Implicit cognition, emotion, and meta-cognitive control. Mind Soc. **11**(1), 107–119 (2012) (special Issue on Dual Processes Theories of Thinking (ed. by D. Over, L. Macchi, R. Viale))
53. R. Sun, I. Naveh, Simulating organizational decision-making using a cognitively realistic agent model. J. Artif. Societies Soc. Simul. **7**(3), (2004)
54. R. Sun, I. Naveh, Social institution, cognition, and survival: A cognitive-social simulation. Mind Soc. **6**(2), 115–142 (2007)

55. R. Sun, P. Slusarz, C. Terry, The interaction of the explicit and the implicit in skill learning: A dual-process approach. Psychol. Rev. **112**(1), 159–192 (2005)
56. R. Sun, N. Wilson, A model of personality should be a cognitive architecture itself. Cogn. Syst. Res. **29–30**, 1–30 (2014)
57. R. Sun, N. Wilson, M. Lynch, Emotion: A unified mechanistic interpretation from a cognitive architecture. Cogn. Comput. **8**(1), 1–14 (2016)
58. F.M. Toates, *Motivational Systems* (Cambridge University Press, Cambridge, UK, 1986)
59. E.C. Tolman, *Purposive Behavior in Animals and Men* (Century, New York, 1932)
60. Toby Tyrrell. *Computational mechanisms for action selection*. PhD thesis, Oxford University, Oxford, UK, 1993
61. C.J.C.H. Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge England, 1989
62. B. Weiner, *Human Motivation: Metaphors, Theories, and Research* (Sage, Newbury Park, CA, 1992)
63. N. Wilson, *Towards a psychologically plausible comprehensive computational theory of emotion*. PhD thesis, Cognitive Sciences Department, Rensselaer Polytechnic Institute, Troy, NY, 2012
64. N. Wilson, R. Sun, Coping with bullying: a computational emotion-theoretic account. *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (Quebec City, Quebec, Canada, 2014), pp. 3119–3124 (Cognitive Science Society, Austin, Texas)
65. N. Wilson, R. Sun, R. Mathews, A motivationally based computational interpretation of social anxiety induced stereotype bias. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (Lawrence Erlbaum Associates, Mahwah, NJ, 2010), pp. 1750–1755
66. N. Wilson, R. Sun, R.C. Mathews, A motivationally-based simulation of performance degradation under pressure. Neural Netw. **22**(5), 502–508 (2009)
67. I. Wright, A. Sloman, Minder 1: An implementation of a protoemotional agent architecture. Technical Report CSRP-97-1, University of Birmingham, School of Computer Science, 1997