AI & SOCIETY (2020) 35:17–28
https://doi.org/10.1007/s00146-017-0775-7

ORIGINAL ARTICLE

Potential of full human–machine symbiosis through truly intelligent cognitive systems

Ron Sun[1]

Received: 10 September 2017 / Accepted: 17 November 2017 / Published online: 28 November 2017

Abstract
It is highly likely that, to achieve full human–machine symbiosis, truly intelligent cognitive systems—human-like (or even beyond)—may have to be developed first. Such systems should not only be capable of performing human-like thinking, reasoning, and problem solving, but also be capable of displaying human-like motivation, emotion, and personality. In this opinion article, I will argue that such systems are indeed possible and needed to achieve true and full symbiosis with humans. A computational cognitive architecture (named Clarion) is used in this article to illustrate, in a preliminary way, what can be achieved in this regard. It is shown that Clarion involves complex structures, representations, and mechanisms, and is capable of capturing human cognitive performance (including skills, reasoning, memory, and so on) as well as human motivation, emotion, personality, and other relevant aspects. It is further argued that the cognitive architecture can enable and facilitate true human–machine symbiosis.

Keywords Cognitive architecture · Emotion · Motivation · Personality · Symbiosis

1 Introduction

It was predicted, a long time ago, that "in not too many years, human brains and computing machines will be coupled together very tightly and the resulting partnership will think as no human brain has ever thought…." (Licklider 1960). Further, Licklider (1960) estimated that "it would be 1980 before developments in artificial intelligence make it possible for machines alone to do much thinking or problem solving…. That would leave, say, five years to develop man–computer symbiosis and 15 years to use it". In so predicting, he was, of course, overly optimistic. He was also solely focused on technical capabilities of computing machines for performing reasoning or problem solving. What was neglected is the possibility that, to achieve full human–machine symbiosis, truly intelligent cognitive machines may have to be developed first. That is, true and full symbiosis between humans and machines may require that truly human-like characteristics be developed in intelligent machines. Such machines not only should be capable of human-like thinking, reasoning, and problem solving, but also should be capable of displaying human-like motivation, emotion, and personality, among other things.

In this short opinion article, I will argue that not only intelligent machines that are highly skilled and knowledgeable and capable of human-like learning and reasoning are important, but it is also important that they are capable of human-like personality, emotion, motivation, and so on in addition to being cognitively intelligent in the narrow sense, in order to be truly symbiotic with humans. I will argue that this is indeed possible, by discussing and utilizing work from cognitive science and computational psychology (Sun 2008).

In the remainder of this article, first, why human-like intelligent machines are needed is discussed from several different perspectives. Then, an approach towards developing such machines is described, in the form of a comprehensive computational cognitive architecture. Such a cognitive architecture will be capable of capturing human cognitive performance (including skills, reasoning, memory, and so on), as well as capturing human motivation, emotion, personality, and other relevant aspects, which in turn facilitate true human–machine symbiosis. Some concluding remarks then complete the article.

✉ Ron Sun
rsun@rpi.edu

[1] Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

2 Why are human-like intelligent machines needed?

Human-like intelligent machines are needed because of potentially tremendous benefits that they may confer. These potential benefits may include those stemming from their being full partners to humans and those stemming from their serving as useful instruments for humans.

As a partner, a truly human-like intelligent machine (or a cognitive machine) is more likely to interact and cooperate well with humans, compared with machines that lack sufficient human-like characteristics because:

- It will be easier for such a machine to communicate with humans, not just through language, but also through other explicit or implicit means of communication.
- It is likely that such a machine can be more easily understood by humans (and, conversely, humans can be more easily understood by such a machine), because humans, with their readily available mental models and "theory of mind" (either innate, or formed through socialization during individual development), can more easily interpret the behavior of such a machine.
- It is more likely that mutual trust can be established between humans and such a machine because of relatively similar characteristics.

And so on. Thus, such systems facilitate the interaction and cooperation between humans and machines because of the relatively high degree of similarity between humans and sufficiently human-like intelligent machines.

For example, let us look specifically into trust between humans and machines. Humans have self-determined and intrinsic motivations and thus are capable of autonomous choice of action in accordance with these motivations. These motivations in humans include not only power, achievement, and other individualistic tendencies, but also adherence to social norms, affiliation with other individuals, and other tendencies related to social cooperation and interdependence (Sun 2009; see details in Sect. 3). These motives are results of evolution over a long period of human prehistory in the context of the struggles to survive by social groups. Consequently, real social trust is the trust among such similarly motivated individuals. Limited, simpler forms of "trust" that one typically places (or does not place) on currently available machines such as self-driving automobiles or robotic vacuum cleaners should not be construed as real trust, or full trust, and is likely not going to be sufficient for future full human–machine symbiosis.

To achieve real trust, we need to delve into natural human tendencies to trust other individuals who have intrinsic motivations that are similar to theirs (Sun 2018). Humans did develop such tendencies, necessitated by their need for survival, evolved during their collective struggles to survive. Such trust may start from predictability of behavior, as a result of similarly endowed (innate or acquired) motives. Understanding others' motivations 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, truly autonomous machines may emerge that may be given our real and full trust.

Some claimed that a machine could not be given our full trust because it never had experiences sufficiently similar to what we had such as losing a parent or suffering from a heart attack. Such a claim would be equivalent to saying that a certain young person could not be given full trust simply because that young person never had such an experience as losing a parent or suffering from a heart attack (cf. Nagel 1974). What is important here, in our view, is similarly endowed intrinsic motives (and other psychological processes; Sun 2009).

For another example, let us look into teaming of humans and autonomous machines, for example, in the form of human-robot teams. Evidently, when a machine understands human motivation and emotion, it works better with humans (Sun 2006). If it can anticipate what a human will need and will do, it can provide better assistance. Furthermore, if it can appreciate, for example, the frustration that a human feels, then it can help to find solutions. If it can understand the anger that a human feels, then it may provide proper counsel. And so on. All this is contingent upon its understanding human motivation and emotion, or better yet, having sufficiently similar, human-like motivation and emotion. When machines have human-like motivation and emotion, they can truly be partners to humans.

Furthermore, beyond being a capable full partner, such a machine may also be employed as a useful instrument. For instance, it can be used for better monitoring and regulating human behavior and performance (including human learning), through its understanding of human behavior, utilizing its characteristics of human-like inner working. Furthermore, it should also be capable of monitoring, recognizing, and regulating people's emotion, motivation, and so on, so that it can help with accomplishing relevant tasks in various complex or difficult situations. This can be better accomplished when the machine in question has sufficiently human-like motivation, emotion, and so on, as well as sufficient reflective and inferential capabilities. These possibilities may exist separately from and in addition to the scenario of human–machine cooperation as full partners or teammates (as discussed earlier). Many other possibilities exist as well, including augmented cognitive systems, cognitive-cyber symbiotic systems, and so on, in which human-like characteristics in terms of cognition, motivation, emotion, and so on can also be helpful.

However, before these possibilities can materialize, a better understanding of the human mind itself is needed, especially a better understanding in a computational form. A better computational understanding of the human mind can lead to truly human-like intelligent cognitive machines (Sun 2008).

3 Towards human-like intelligent machines for true symbiosis

3.1 Cognitive architecture in cognitive science

To demonstrate possibilities of working towards truly human-like intelligent machines that may lead to full human–machine symbiosis, I will describe some research that aims to develop psychologically realistic computational cognitive architectures, out of the fields of cognitive science and computational psychology, and may serve as the basis for truly intelligent cognitive machines.

Cognitive science is the interdisciplinary study of the mind in terms of mechanisms and processes that constitute the mind. The fundamental assumption of cognitive science is that the human mind may be understood in terms of representational structures and computational procedures that operate on those structures. Although the roots of cognitive science can be traced back to much earlier times, its modern beginning started as an intellectual movement around the mid-century (the cognitive revolution). In particular, Newell and Simon's early computational work in the 1960s and 1970s has been seminal (see, e.g., Newell 1990).

Among other methodologies, computational modeling (computational psychology) is an extremely important aspect in cognitive science. Computational models in cognitive science are essentially mechanistic, process-oriented theories (for the most part). That is, they are mostly aimed at answering the questions of how human performance comes about, by what psychological structures, mechanisms, and processes, and in what ways. The key to understanding psychological phenomena is often in fine details, which computational modelling can illuminate (Newell 1990; Sun 2007). It embodies specific descriptions of psychological processes in computer algorithms and programs. That is, it imputes computational processes onto psychological functions, and thereby it produces runnable computational models. Detailed simulations are conducted based on the computational models (see, e.g., Rumelhart et al. 1986). Computational models provide algorithmic specificity: detailed, exactly specified, and carefully worked-out steps, arranged in precise and yet flexible sequences. Thus, they provide clarity and precision (Sun 2008).

In particular, a computational cognitive architecture, as commonly termed in cognitive science, is a broadly scoped, domain-generic cognitive-psychological model, implemented computationally, capturing the essential structures, mechanisms, and processes of the mind, to be used for a broad, multiple-level, multiple-domain analysis of behavior (e.g., through its instantiation into more detailed computational models or as a general framework; Newell 1990; Sun 2007). A cognitive architecture can be important to cognitive science: It provides concrete computational scaffolding for more detailed modeling and exploration of cognitive-psychological phenomena and data, through specifying essential computational structures and mechanisms.

Moreover, broad functionalities commonly found in cognitive architectures are even more important (Newell 1990). The human mind needs to deal with all of its functionalities: perception, categorization, memory, decision-making, reasoning, problem solving, communication, action, learning, metacognition, motivation, and so on. The need for generic models capable of these broad functionalities arises because of the need to avoid the fragmentation often resulting from narrowly scoped research.

Thus, developing cognitive architectures is an important endeavor in cognitive science and computational psychology (Sun 2008). It is of fundamental importance in advancing the understanding of the human mind (Sun 2002, 2016).

Cognitive architectures can also facilitate the building of truly human-like intelligent machines. In relation to building intelligent systems, a cognitive architecture may provide underlying infrastructures because it may include a variety of capabilities, modules, and mechanisms that a human-like intelligent system needs. On that basis, intelligent systems may be more readily developed. A cognitive architecture carries with it theories of psychology and understanding of intelligence gained from exploring the human mind. In a way, cognitive architectures reverse-engineer the human mind. Therefore, the development of intelligent systems on that basis may be more psychologically grounded and more psychologically realistic, which may be useful towards achieving true human–machine symbiosis (as discussed earlier).

Existing cognitive architectures include ACT-R, Soar, Clarion, and a number of others (see, e.g., the chapter on cognitive architectures in Sun 2008 for a review). Among them, in particular, Clarion is a generic and comprehensive computational cognitive architecture aimed to capture, explain, and simulate a very wide variety of cognitive-psychological phenomena within its unified framework, thus leading to unified explanations of psychological phenomena (as advocated by, e.g., Newell 1990). Two points stand out:

- Clarion is more comprehensive in scope than most other cognitive architectures in existence today;
- Clarion is psychologically realistic to the extent that it has been validated through simulating and explaining a very wide variety of psychological tasks, data, and phenomena.

The exact extent of psychological phenomena that have been captured and explained within its framework has been discussed in detail in prior publications (see, e.g., Sun 2002, 2016; Sun et al. 2001, 2005; Helie and Sun 2010; Bretz and Sun 2017). It is not unreasonable to say that Clarion constitutes an initial version of a (relatively) comprehensive theory of the mind.

3.2 A review of the Clarion cognitive architecture

3.2.1 Overview of Clarion

Clarion provides structural and algorithmic specifications of a wide range of psychological processes. Only a sketch of Clarion can be presented below; the vast majority of technical details are omitted due to length considerations. See Fig. 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 (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 (Sun et al. 2005). The role of the non-action-centered subsystem is to maintain and utilize declarative knowledge (Helie and Sun 2010). The role of the motivational subsystem is to provide underlying motivations for perception, action, and cognition (in terms of providing impetus and feedback; Sun 2009; Merrick and Maher 2009; Baldassarre and Mirolli 2013). The role of the metacognitive subsystem is to monitor, direct, and modify the operations of the other subsystems dynamically (e.g., Reder 1996; Sun et al. 2006).

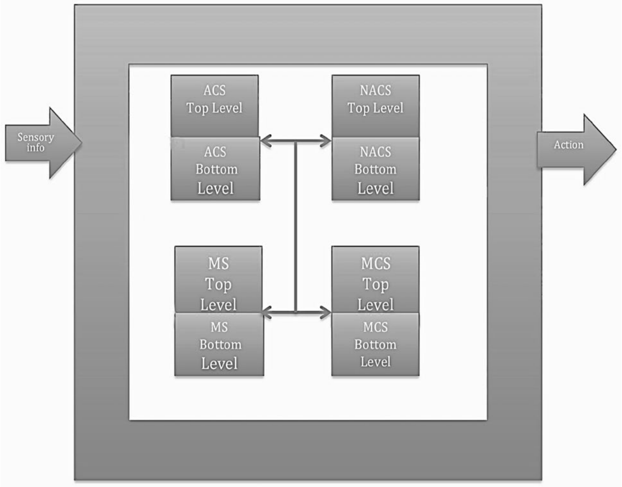

Fig. 1 The Clarion cognitive architecture. The subsystems of Clarion are shown. The major information flows are shown with arrows. ACS stands for the action-centered subsystem. NACS stands for the non-action-centered subsystem. MS stands for the motivational subsystem. MCS stands for the metacognitive subsystem

Each of these interacting subsystems consists of two "levels" of representations (i.e., a dual-representational structure, as theoretically posited in Sun 2002). Generally speaking, in each subsystem, the "top level" encodes explicit knowledge (using symbolic/localist representations) and the "bottom level" encodes implicit knowledge (using distributed representations; Rumelhart et al. 1986). Roughly speaking, explicit knowledge is directly consciously accessible (Reber 1989), 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 Sun (2002, 2016) for further details.

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 (Sun et al. 2001, 2005).

Existing theories tend to confuse implicit and explicit processes; hence the "perplexing complexity" (Smillie et al. 2006). In contrast, Clarion generally separates 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, e.g., Sun et al. 2001, 2005; Helie and Sun 2010; Bretz and Sun 2017).

Furthermore, Clarion accounts for basic human motives, which provide the underlying basis for behavior. This emphasis on human motivation 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.

3.2.2 The action-centered subsystem

The ACS captures the process of human action selection: 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 selected action is then performed. The action changes the world in some way. Comparing the changed state of the world with the previous state, the individual learns. The cycle then repeats itself.

In this subsystem, the bottom level consists of "action neural networks" encoding implicit procedural knowledge (involving distributed representations; Rumelhart et al. 1986), and the top level consists of "action rules" encoding explicit procedural knowledge (using symbolic/localist representations; Sun 2002).

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:

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

where $p\ (a|x)$ is the probability of selecting action a, τ (temperature) controls the degree of randomness of action selection, and i ranges over all possible actions. (This is known as Luce's choice axiom; Watkins 1989.)

For learning implicit knowledge at the bottom level (i.e., the Q values), the Q learning algorithm (Watkins 1989), 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 Watkins 1989; Sun et al. 2001).

For learning explicit knowledge at the top level (i.e., action rules), a variety of algorithms may be applied, including the rule-extraction-refinement (RER) algorithm for a "bottom–up" learning process that relies on implicit knowledge from the bottom level to learn explicit knowledge at the top level (Sun et al. 2001). In the reverse direction, "top–down" learning can also occur (Sun 2016).

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 (Sun et al. 2001). The selection probabilities may be variable, determined by the metacognitive subsystem (by its processing mode module; more later).

3.2.3 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; Rumelhart et al. 1986).

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 (Sun 2016).

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 implicit knowledge at the bottom level, or to assimilate explicit knowledge of the top level into 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, e.g., Helie and Sun 2010; Sun 2016). Together they enable the NACS to capture much of human everyday reasoning (Sun 2016).

3.2.4 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 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. 2. For justifications, see Sun (2009).

At the bottom level of the MS, primary drives are motives that are 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 (Murray 1938; Reiss 2004; Sun 2009). See Table 1 for their descriptions.

These primary drives have been justified. Briefly, this set of drives bears close relationships to Murray's needs (1938), Reiss's motives (2004), and so on. The prior

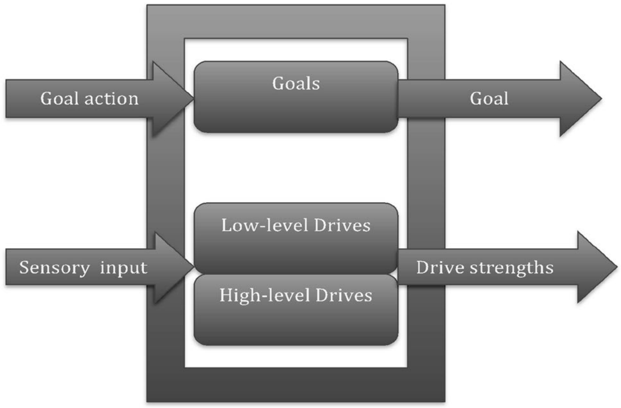

Fig. 2 The basic structure of the motivational subsystem

Table 2 Approach versus avoidance primary drives

Approach drives	Avoidance drives	Both
Food	Sleep	Affiliation and belongingness
Water	Avoiding danger	Similance
Reproduction	Avoiding unpleasant stimuli	Deference
Nurturance	Honor	Autonomy
Curiosity	Conservation	Fairness
Dominance and power		
Recognition and achievement		

justifications of these frameworks can be applied, to a significant extent, to this set of drives as well (see Murray 1938; Reiss 2004; Sun 2009). On the basis of primary drives, secondary (derived) drives may be acquired.

Some of these primary drives are approach-oriented, while some others are avoidance-oriented. This distinction has been argued by many (e.g., Clark and Watson 1999; Gray and McNaughton 2000; Smillie et al. 2006). The approach system is sensitive to cue 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 2 for this division of drives.

The processing of these drives within the bottom level of the MS involves a number of modules (Sun 2016). In particular, the core drive module determines drive strengths (using neural networks) based roughly on:

$$\mathrm{ds}_d = \mathrm{gain}_d \times \mathrm{stimulus}_d \times \mathrm{deficit}_d + \mathrm{baseline}_d$$

where ds_d is the strength (activation) of drive d, gain_d is the gain for drive d, $\mathrm{stimulus}_d$ is a value representing how pertinent the current situation is to drive d, $\mathrm{deficit}_d$ indicates the perceived deficit in relation to drive d (which represents an individual's intrinsic inclination toward activating drive d), and $\mathrm{baseline}_d$ is the baseline strength of drive d. The justifications for this can be found in the literature (e.g., Tyrell 1993; Sun 2009, 2016).

Motivational adaptation (learning) is also possible and has been tackled (e.g., Sun and Wilson 2014). In addition, new drives ("derived drives") may be acquired. They may be gradually acquired through some kind of "conditioning"

Table 1 Primary drives within Clarion

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 and belongingness	The drive to associate with other individuals and to be part of social groups
Dominance and power	The drive to have power over other individuals
Recognition and 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

or may be externally set through externally provided instructions, on the basis of primary drives.

3.2.5 The metacognitive subsystem

Metacognition refers to active monitoring and regulation of one's own psychological processes (Reder 1996). In Clarion, the MCS is closely tied to the MS. The MCS monitors, controls, and regulates other processes. 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 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. 3.

For instance, the goal module selects goals to pursue (for the ACS). To select a new goal, it first determines goal strengths, based on information from the MS (e.g., drive strengths). Then, a new goal is stochastically selected on the basis of the goal strengths (using a Boltzmann distribution). For arguments in support of goal setting on the basis of implicit motives (i.e., drives), see, for example, Tolman (1932). In the simplest case, the following calculation is performed:

$$\mathrm{gs}_g = \sum_{d=1}^{n} \mathrm{relevance}_{d,s\to g} \times \mathrm{ds}_d$$

where gs_g is the strength of goal g, $\mathrm{relevance}_{d,s\to g}$ is a measure of how relevant drive d is to goal g with regard to the 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 sake of integrating outcomes from the two levels of the ACS (see the discussion of

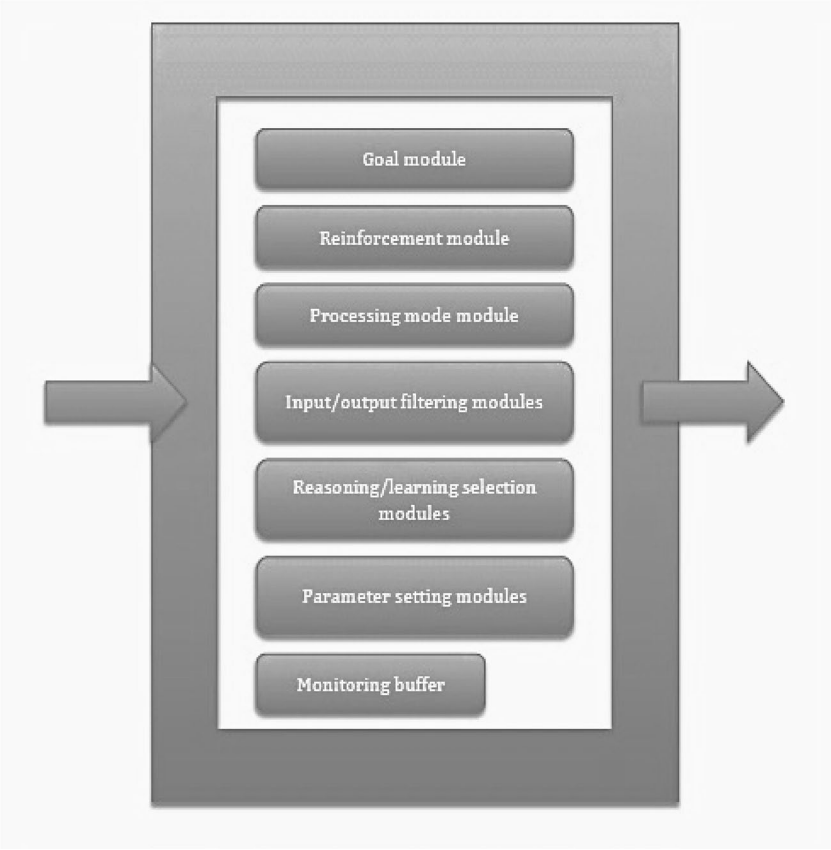
Goal module
Reinforcement module
Processing mode module
Input/output filtering modules
Reasoning/learning selection modules
Parameter setting modules
Monitoring buffer

Fig. 3 The main modules within the metacognitive subsystem

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 Sun 2016 for details). However, these probabilities may be modulated multiplicatively by another parameter: the strength of avoidance-oriented drives (which corresponds to "anxiety"; Wilson et al. 2009).

4 Capturing truly human-like characteristics

Clarion has been successful in computationally capturing, modeling, simulating, and explaining a wide variety of psychological data and phenomena. In the first subsection below, I will briefly summarize modeling procedural and declarative processes. In the second subsection, I will summarize models involving motivational and metacognitive processes, covering personality, emotion, motivation, and so on. In the third subsection, these results are brought to bear on enabling true human–machine symbiosis.

Note that, while accounting for various psychological tasks, data, and phenomena, Clarion provides explanations that shed significant new theoretical light on underlying psychological processes. See, for example, Sun et al. (2001), Sun et al. (2005), Helie and Sun (2010), and Bretz and Sun (2017) for various examples.

4.1 Capturing procedural and declarative processes in Clarion

4.1.1 Capturing procedural skills and knowledge

Many tasks involving procedural processes have been captured and simulated using Clarion. For example, a number of skill-learning tasks have been simulated that span the spectrum ranging from simple reactive skills to complex cognitive skills. These simulated tasks include commonly used psychological tasks such as serial reaction time tasks, artificial grammar-learning tasks, dynamic process-control tasks, alphabetical arithmetic tasks, and so on (Sun et al. 2005; Sun 2002).

Furthermore, it should be pointed out that work has been done in modeling complex and realistic skill tasks that involve complex and realistic sequential decision-making, beyond typical laboratory tasks. For example, a complex minefield navigation task was tackled (see Sun et al. 2001). More recently, work is being done on tackling human intelligence tests, such as Raven's Progressive Matrices, to better understand the nature of human intelligence (Mekik et al. 2017).

4.1.2 Capturing declarative knowledge and reasoning

Many tasks involving declarative processes have also been modeled and simulated. They include various reasoning tasks, involving not only explicit reasoning but also intuition and insight. For example, Sun and Zhang (2006) showed how patterns demonstrated by humans in these reasoning tasks might be captured in Clarion. Helie and Sun (2010) showed in depth how intuition developed and emerged as insight through simulating a variety of relevant human data.

In addition, Helie and Sun (2014) and Sun and Helie (2013) examined a large number of psychological "laws" (regularities) concerning human reasoning and memory that Clarion is able to account for, and explained in a principled way how Clarion accounts for them.

4.2 Capturing motivational and metacognitive processes in Clarion

Many other kinds of tasks that are not usually dealt with by other cognitive architectures have been tackled in Clarion. In particular, tasks involving motivational and metacognitive processes have been addressed. On that basis, social interactions have been modeled and large-scale social simulations have been carried out.

Below, I will describe the modeling of motivation, emotion, and personality. For other motivational or metacognitive simulations, see Sun et al. (2006), Wilson et al. (2009), Bretz and Sun (2017), and so on.

4.2.1 Capturing motivation

On the basis of the mechanisms described in Sect. 3, Clarion can account for many psychological phenomena related to human motivation. For instance, Lambert et al. (2003) showed that in socially stressful situations, social stereotyping was more pronounced. They used the task of recognition of tools versus guns, when primed by black or white faces. The results showed that, in socially stressful situations, when paired with a black face, tools were much more likely to be mistaken as guns. This phenomenon has been captured, explained, and simulated using Clarion. When certain avoidance-oriented drive strengths become very high within the MS (e.g., as a result of stressful situations), the processing within the ACS becomes more implicit, as determined by the MCS on the basis of drive strength levels (see Sect. 3). The implicit processing within the ACS is more susceptible to stereotyping effects due to the nature of implicit learning. The simulation using Clarion captured the corresponding human data (Wilson et al. 2009) and provided a detailed, mechanistic, and process-based explanation for the data.

Likewise, skilled performance may deteriorate when individuals are under pressure. In terms of mathematical skills, Beilock et al. (2004) showed that performance worsened when pressure was high. They used a modular arithmetic problem set (of the form $A = B$ mod C), and tested participants either under pressure or not. The result showed significant differences with versus without pressure. This task has been captured using Clarion, which provided mechanistic, process-based explanations. When certain avoidance-oriented drive strengths within the MS are high (e.g., as a result of pressure), processing within the ACS becomes very implicit (controlled by the MCS on the basis of drive strength levels). Overly implicit processing leads to worsened performance (Sun et al. 2005). The simulation captured the corresponding human data (Wilson et al. 2009).

Note that, what is most important with regard to motivation is the link from an internal need, which is triggered by situational cues, to the selection of an action, as well as to internal parameters relevant for selecting an action. It is not particularly important whether the entity embodying this process is a machine or a human. What matters are the internal mechanisms by which the need arises and leads to corresponding actions (given relevant cues), which can be readily described computationally (Sun 2009). For instance, "honor" involves adherence to social norms and codes in relevant situations (see Sect. 3). A robot can have this internal need and thus perform appropriate actions accordingly in relevant circumstances (provided that it learned relevant norms and codes), by the afore-mentioned mechanisms.

4.2.2 Capturing personality

On the basis of the models of motivation and metacognition (see Sect. 3), human personality may be accounted for as well. The Clarion personality model is first based on drives within the MS. On that basis, goal setting (by the MCS) and action selection (by the ACS) take place. Individual differences may be accounted for (for the most part) by the differences in relative drive strengths in different situations by different individuals. Individual differences in terms of drive strengths are consequently reflected in the resulting goals, major cognitive parameters, and action selection on that basis. Personality types, besides being mapped onto drive activation parameters, are also mapped onto other parameters involving other mechanisms and processes (although to a lesser degree). For instance, within Clarion, personality may involve parameters within the ACS, the NACS, the MS, and the MCS. Therefore, personality is the result of complex interactions among a large set of mechanisms. This approach can be justified from a variety of perspectives; see Sun and Wilson (2014) for details.

Various tests show that the Clarion personality model is capable of demonstrating stable personality traits and at the same time showing sufficient variability of behaviors in response to different situations (Sun and Wilson 2014). It maps onto the well-known Big Five personality structures (cf. Read et al. 2010).

This model has been used to simulate and explain a variety of relevant human data. For instance, Moskowitz et al. (1994) examined the influence of social role/status on interpersonal behavior and hypothesized that social role/status would have various effects on behavior. Subjects were expected to behave more submissively, for example, when interacting with a boss versus a coworker or a subordinate. Event contingent recording was used to gather data and the data confirmed these effects. The Clarion simulation with the personality model above captured all the major effects exhibited within the human data. Various other simulations of human data have also been carried out (see Sun and Wilson 2014).

4.2.3 Capturing emotion

According to Clarion, emotion is the result of many processes throughout a system. Its may involve physiological states, action readiness, physical and mental actions, motivational processes, evaluation and attribution processes, metacognitive processes, as well as decision-making and reasoning of various kinds. According to Clarion, emotion is the sum total of all of the above in particular circumstances (Sun et al. 2016).

Specifically, in Clarion, emotion is captured by a multitude of processes involving the ACS (for action), the NACS (for evaluation), the MS (for motivation), and the MCS (for metacognitive regulation). In particular, emotion is closely related to the MS. Smillie et al. (2006), Carver and Scheier (1998), and Ortony et al. (1988) stressed the importance of motivation and expectation in emotion. For instance, it has been hypothesized within Clarion that the emotion of elation may be related to positive reward (including "unexpected" positive reward) and also, to a lesser extent, "expectation" of positive reward. Computationally, the intensity of elation may be in part a function of drive strengths of approach-oriented drives in the MS, which (in part) determine reward. For another instance, the emotion of anxiety may be related to "expectation" of negative reward. The intensity of anxiety may be in part a function of strengths of avoidance-oriented drives in the MS. In this regard, Smillie et al. (2006) discussed the link between the avoidance system and anxiety; see also Carver and Scheier (1998, p. 92). Furthermore, emotion is closely related to the ACS, because it is closely tied to action. Frijda (1986), for example, indicated the importance of "action readiness" in emotional experience.

Emotional processing mainly occurs in the bottom level of Clarion in various subsystems (Sun et al. 2016); that is, emotional processing is mostly implicit (although not all implicit processes are emotional; Damasio 1994). Explicit processes may also have some role in emotion, for example, through affecting decisions of the bottom level or through explicit reasoning for "cognitive appraisal" (Frijda 1986), although they are not the main locus of emotion according to Clarion. Various simulations of human data have been carried out accordingly.

4.3 Enabling and facilitating symbiosis

Overall, as discussed above, the Clarion project addresses the understanding and modeling of essential procedural, declarative, motivational, metacognitive, and other processes in humans and in highly human-like systems. Clarion includes representations, structures, and mechanisms necessary for a comprehensive computational model of the human mind. It constitutes a requisite step towards making computational cognitive architectures more realistic models of the mind, taking into consideration all of its complexity and intricacy. It is thus also highly relevant to developing truly autonomous computational agents or machines capable of functioning autonomously in complex environments working with their human counterparts. Note that, what I emphasize here is developing highly human-like, fully autonomous agents or machines; true and full symbiosis of such machines with humans is possible and will likely be needed in the future (as argued earlier in Sect. 2).

Based on work so far, there are reasons to be confident that all the important aspects in this regard can be successfully tackled (at least to a sufficient extent). These aspects have been tackled within a unified framework (namely Clarion), so there is indeed some convergence. Note that, our focus has been on models that capture essential human-like motivation, emotion, personality, intuition, reasoning, skill, memory, and so on, especially in the context of human everyday activities (Heidegger 1927). This focus is substantially different from those of the current AI community; in particular, it addresses detailed psychological processes, on the firm basis of empirical work in psychology (as well as a number of other empirical fields). Putting these pieces together, an agent model could indeed be human-like.

To address this point further, take the examples of trust between machines and humans and intrinsic motivation in humans and machines, as mentioned in Sect. 2. First, "free will" in humans or human-like agents and machines implies self-determined, intrinsic motivation and autonomous choice of action in accordance with intrinsic motivation (Sun 2018). It implies capacities by agents to make decisions autonomously at their own discretion. However, their decision space is actually also well bounded and shaped by social ties, socially oriented needs, social rules and norms, and interdependencies among individuals. So intrinsic motivations in these regards (as discussed in Sect. 3) are extremely important also.

It may be worthwhile to address briefly the specific notion of "free will" used here, by which I meant "intrinsic motivation and autonomous choice of action in accordance with intrinsic motivation". Without getting into extensive philosophical treatments of this topic (including expositions of various compatibilist views; e.g., Hume 1765), the above definition implies (1) existence of internal needs/motives, and (2) choice of action in accordance with such needs/motives (dealing with their convergence or divergence and situational factors), (3) by specific, describable mental mechanisms and processes. All of these aspects have been addressed in Clarion (Sun 2016). (Note that, qualia or phenomenal consciousness in this regard are not addressed in this work.)

Second, let us turn to trust. Trust is sometimes defined simply as the confidence (the estimated likelihood) that an autonomous agent will help to achieve another autonomous agent's goals in some types of situations (Abbass et al. 2017). However, as discussed before, trust is more appropriately viewed here as natural human tendencies to understand, to empathize, and to rely on other individuals with intrinsic motivations that are similar to their own (Sun 2018). In other words, from our perspective, limited trust, superficial trust, or forced trust are not construed as real trust; real trust is necessarily deeper.

There has been a great deal of work on the structure of human motivation, so we already know a great deal about it (Murray 1938; Reiss 2004). As discussed earlier, implicit drives, as well as explicit goals, have been structured into Clarion (Sun 2009, 2016). On the basis of drives, explicit goals may be generated, which guide action selection. Such understanding has led to computational simulations including Wilson et al. (2009) and Bretz and Sun (2017). For multi-agent simulations concerning motivation, see Sun and Fleischer (2012). Other motivation-related models were discussed earlier. The motivational representations and mechanisms and their resulting dynamics help to make a computational cognitive architecture functioning in a more psychologically realistic way. More importantly, using such a model, understanding others' motivation becomes possible. Understanding others' motivation leads to predictability of their behavior, which in turn leads to real trust as a result of evolved human nature to place confidence on other individuals whose motivation can be understood (including invoking relevant affective or emotional processes; Sun et al. 2016).

Following this path, through understanding and exploiting natural human tendencies, truly autonomous agents, robots, and machines may emerge in the future that may be given our real and full trust, leading to true symbiosis with humans. When humans place their trust on them and are in turn rewarded with greater reliability, simplicity, safety, and other useful features in a somewhat consistent way, trust will grow and take hold. True trust on and true symbiosis with autonomous machines can ultimately be achieved (hopefully in the not-so-distant future).

Beyond this example above, there are, of course, many other examples, cases, and scenarios where highly human-like characteristics in machines are needed to achieve true human–machine symbiosis (such as those enumerated in Sect. 2).

To further develop such an approach for exploring possibilities of true human–machine symbiosis, future challenges abound, which include applying the framework sketched above to the building of next-generation intelligent systems that are much more human-like and generate human-like behavior with more robustness, flexibility, and versatility that mesh well with human behavior, which will be an ultimate test of the usefulness and feasibility of this framework. Another significant challenge that underlies this enterprise is to further validate, in a careful and detailed way, through empirical work (especially empirical psychological work), this framework and its theoretical implications (separate from building practical intelligent systems). Many more experiments, simulations, and other tests will be needed and shall be pursued in the future.

One possible criticism is that this approach may be overly optimistic about the prospect for human–machine cooperation and symbiosis; there could be a downside to fully autonomous systems: If they are truly independent, what guarantees that their motivations and actions will be beneficial to or at least compatible with humans'? It is worth pointing out that, by and large, humans managed to stay largely cooperative, if we take a long-term and broad view on this, despite the existence of antagonism, competition, and conflict. If we take a long-term and broad view, we have reasons to hope that, if humans can manage to cooperate among themselves, then humans and truly human-like systems can also learn to cooperate with each other (since they are so similar). At least, truly human-like systems will be more predictable from a human perspective than arbitrarily engineered systems whose long-term outcomes we cannot possibly foresee in all circumstances.

Whether or not we will eventually achieve such results is an empirical question that is of tremendous consequences. There is no a priori reason to believe in one way or the other at this point, although there have been many philosophical speculations. Instead of pre-maturely declaring success or failure, we need to work on the relevant aspects and issues incrementally and at the same time keep in mind the big picture.

5 Concluding remarks

In this article, I have argued that truly intelligent cognitive machines—capable of human-like motivation, emotion, and personality, highly skilled and knowledgeable, and performing human-like reasoning and learning—are important for achieving full and true symbiosis with humans, and they may be possible to develop.

To achieve full human–machine symbiosis in the future, these aspects need to be further explored and developed, beyond developing computational models around usual topics such as deep learning, reinforcement learning, and so on. It is necessary that we go beyond these popular techniques for artificial/computational intelligence and explore much further. In particular, we need to look into cognitive science and psychologically realistic computational cognitive architectures resulting from it, which have the potential for leading up to truly intelligent, truly human-like machines, which in turn may help to achieve full human–machine symbiosis. In this article, the Clarion cognitive architecture has been used as an example to show, in an admittedly preliminary way, how it can capture human-like cognitive capabilities as well as human-like motivation, emotion, personality, and so on. Such a model is a necessary step towards being able to build truly human-like machines that are capable of fully cooperating with human beings in a human-like way, for the benefit of all involved.

Acknowledgements This work was supported in part by the ARI Grant W911NF-17-1-0236. Thanks are due to the reviewers, who provided useful suggestions.

References

Abbass H, Scholz J, Reid D (eds) (2017) Foundations of trusted autonomy. Springer, Berlin

Baldassarre G, Mirolli M (2013) Intrinsically motivated learning in natural and artificial systems. Springer, Berlin

Beilock S, Kulp C, Holt L, Carr T (2004) More on the fragility of performance: choking under pressure in mathematical problem solving. J Exp Psychol Gen 133(4):584–600

Bretz S, Sun R (2017) Two models of moral judgment. Cogn Sci. https://doi.org/10.1111/cogs.12517

Carver C, Scheier M (1998) On the self-regulation of behavior. Cambridge University Press, Cambridge

Clark LA, Watson D (1999) Temperament: a new paradigm for trait psychology. In: Pervin LA, John OP (eds) Handbook of personality: theory and research, 2nd edn. Guilford Press, New York, pp 399–423

Damasio A (1994) Descartes' error: emotion, reason and the human brain. Grosset/Putnam, New York

Frijda N (1986) The emotions. Cambridge University Press, New York

Gray JA, McNaughton N (2000) The neuropsychology of anxiety: an enquiry into the functions of the septo-hippocampal system, 2nd edn. Oxford University Press, New York

Heidegger M (1927/1962) Being and time. English translation published by Harper and Row, New York

Helie S, Sun R (2010) Incubation, insight, and creative problem solving: a unified theory and a connectionist model. Psychol Rev 117(3):994–1024

Helie S, Sun R (2014) An integrative account of memory and reasoning phenomena. New Ideas Psychol 35:36–52

Hume D (1765/1993). An enquiry concerning human understanding. Hacket Publishing Co., Indianapolis

Lambert A, Payne B, Jacoby L, Shaffer L, Chasteen A, Khan S (2003) Stereotypes as dominant responses: on the "social facilitation" of prejudice in anticipated public contexts. J Pers Soc Psychol 84(2):277–295

Licklider JCR (1960). Man-computer symbiosis. IRE Trans Hum Factors Electron HFE-1:4–11

Mekik CS, Sun R, Dai DY. (2017). Deep learning of Raven's matrices. In: P. Bello (ed.), Proceedings of the fifth annual conference on advances in cognitive systems (ACS 2017), Troy, New York

Merrick E, Maher ML (2009) Motivated reinforcement learning. Springer, Berlin

Moskowitz DS, Suh EJ, Desaulniers J (1994) Situational influences on gender differences in agency and communion. J Pers Soc Psychol 66:753–761

Murray H (1938) Explorations in personality. Oxford University Press, New York

Nagel T (1974) What is it like to be a bat? Philos Rev 83(4):435–450

Newell A (1990) Unified theories of cognition. Harvard University Press, Cambridge

Ortony A, Clore G, Collins A (1988) The cognitive structure of emotions. Cambridge University Press, Cambridge

Read SJ, Monroe BM, Brownstein AL, Yang Y, Chopra G, Miller LC (2010) Virtual personalities II: a neural network model of the structure and dynamics of human personality. Psychol Rev 117:61–92

Reber AS (1989) Implicit learning and tacit knowledge. J Exp Psychol 118(3):219–235

Reder LM (1996) Implicit memory and metacognition. Erlbaum, Mahwah

Reiss S (2004) Multifaceted nature of intrinsic motivation: the theory of 16 basic desires. Rev Gen Psychol 8(3):179–193

Rumelhart DE, McClelland JL, PDP Research Group (1986) Parallel distributed processing. MIT Press, Cambridge

Smillie LD, Pickering AD, Jackson CJ (2006) The new reinforcement sensitivity theory: implications for personality measurement. Personal Soc Psychol Rev 10:320–335

Sun R (2002) Duality of the mind. Lawrence Erlbaum Associates, Mahwah

Sun R (ed) (2006) Cognition and multi-agent interaction: from cognitive modeling to social simulation. Cambridge University Press, New York

Sun R (2007) The importance of cognitive architectures: an analysis based on CLARION. J Exp Theor Artif Intell 19(2):159–193

Sun R (ed) (2008) The Cambridge handbook of computational psychology. Cambridge University Press, New York

Sun R (2009) Motivational representations within a computational cognitive architecture. Cogn Comput 1(1):91–103

Sun R (2016) Anatomy of the mind: exploring psychological mechanisms and processes with the Clarion cognitive architecture. Oxford University Press, New York

Sun R (2018) Intrinsic motivation for truly autonomous agents. In: Abbass H, Scholz J, Reid D (eds) Foundations of trusted autonomy. Springer, Berlin

Sun R, Fleischer P (2012) A cognitive social simulation of tribal survival strategies: The importance of cognitive and motivational factors. J Cogn C 12(3–4):287–321

Sun R, Helie S (2013) Psychologically realistic cognitive agents: taking human cognition seriously. J Exp Theor Artif Intell 25:65–92

Sun R, Wilson N, (2014). A model of personality should be a cognitive architecture itself. Cogn Syst Res 29–30:1–30

Sun R, Zhang X (2006) Accounting for a variety of reasoning data within a cognitive architecture. J Exp Theor Artif Intell 18(2):169–191

Sun R, Merrill E, Peterson T (2001) From implicit skills to explicit knowledge: a bottom–up model of skill learning. Cogn Sci 25(2):203–244

Sun R, Slusarz P, Terry C (2005) The interaction of the explicit and the implicit in skill learning: a dual-process approach. Psychol Rev 112(1):159–192

Sun R, Zhang X, Mathews R (2006) Modeling meta-cognition in a cognitive architecture. Cogn Syst Res 7(4):327–338

Sun R, Wilson N, Lynch M (2016) Emotion: a unified mechanistic interpretation from a cognitive architecture. Cogn Comput 8(1):1–14

Tolman EC (1932) Purposive behavior in animals and men. Century, New York

Tyrell T (1993) Computational mechanisms for action selection. Ph.D. Thesis, Oxford University, Oxford, UK

Watkins C (1989) Learning with delayed rewards. Ph.D. Thesis, Cambridge University, Cambridge, UK

Wilson N, Sun R, Mathews R (2009) A motivationally-based simulation of performance degradation under pressure. Neural Netw 22:502–508