JOURNAL OF EXPERIMENTAL & THEORETICAL ARTIFICIAL INTELLIGENCE, 2018
VOL. 30, NO. 1, 13–37
https://doi.org/10.1080/0952813X.2017.1413141

Taylor & Francis
Taylor & Francis Group

Why is a computational framework for motivational and metacognitive control needed?

Ron Sun

Department of Cognitive Science, Rensselaer Polytechnic Institute, Troy, NY, USA

ABSTRACT

This paper discusses, in the context of computational modelling and simulation of cognition, the relevance of deeper structures in the control of behaviour. Such deeper structures include motivational control of behaviour, which provides underlying causes for actions, and also metacognitive control, which provides higher-order processes for monitoring and regulation. It is argued that such deeper structures are important and thus cannot be ignored in computational cognitive architectures. A general framework based on the Clarion cognitive architecture is outlined that emphasises the interaction amongst action selection, motivation, and metacognition. The upshot is that it is necessary to incorporate all essential processes; short of that, the understanding of cognition can only be incomplete.

ARTICLE HISTORY
Received 20 February 2017
Accepted 11 September 2017

KEYWORDS
Cognitive architecture; motivation; metacognition; implicit cognition; explicit cognition; cognitive modelling

1. Introduction

The motivational and metacognitive underpinnings of human behaviour are often ignored in computational cognitive modelling. During a relatively short laboratory behavioural experiment, motivational issues indeed may not be particularly significant (as such an experiment is often quite different from real-world situations). For example, in playing Tower of Hanoi, participants' motivation may be simply limited to getting the work done in accordance with the instructions. However, if the task goes on long enough, complex interplays of different motivational forces, as one may see in many real-world activities, may show up and will have to be taken into consideration (Dai & Sternberg, 2004; Simon, 1967; Sun, 2009). Such motivations, beside those resulting from basic physiological needs such as hunger and thirst, may also include curiosity, desire to prove one's own capability, desire to conform to a social norm and so on. Metacognitive regulation may likewise be involved.

Most proposed computational cognitive architectures for computational cognitive modelling that are supposed to be psychologically realistic have, more or less, ignored this aspect of human behaviour (mirroring similar neglect in some other lines of work). For example, in ACT-R, a production system based cognitive architecture, goals are set by production rules, which are hand coded into the model (Anderson & Lebiere, 1998). There is no deeper explanation of why a certain goal is set or changed during performance, beside the fact that a production rule sets or changes the goals to help accomplish a task. Hence, ACT-R per se does not shed any new light on motivational aspects of human behaviour.

Production systems are known for their 'brittleness', as discussed by a number of researchers in the past, both in AI and in cognitive science (see, e.g. Hayes-Roth, Waterman, & Lenat, 1983; Sun, 1994). A typical production system has to have precisely specified goals and conditions for determining actions.

CONTACT Ron Sun ✉ rsun@rpi.edu

Otherwise, it would be at a loss as to what to do (Sun, 1994). The same problem is also present in production system-based cognitive architectures. Contrary to the brittleness of these artificial systems, the human mind involves complex interaction among cognitive, motivational and metacognitive processes (Dai & Sternberg, 2004; Weiner, 1992). Possibly as a result of such interactions, humans seem to be able to adapt to different situations rapidly, with more robust, more flexible goal setting and other processes (e.g. Eccles & Wigfield, 2002; Latham & Pinder, 2005).

Given the state of the art, it appears necessary to emphasise motivational and metacognitive processes and how they interact with other (regular) cognitive processes. In particular, we need to examine the cognition-motivation interaction. Essential motivations of an agent arise naturally, prior to deliberative cognition. Such motivations are in fact the foundation of cognition. In a way, cognition has evolved to serve the essential needs and motivations of an agent. In this process, different types of control are present: primary control of actions directly affecting the external environments, and deeper control by motivational and metacognitive mechanisms (Atkinson, 1964; Dai & Sternberg, 2004; Hull, 1951; Toates, 1986; Sun, 2009).

To see the importance of motivational and metacognitive processes, let us examine the case of an agent (e.g. a robot). Without motivational processes, the agent would be literally aimless. It would wonder around the world aimlessly, hardly accomplishing anything. Or it would have to rely on knowledge hand-coded into it, for example, production rules regarding goals and actions, with all the brittleness problems associated with production rules (Sun, 1994), in order to just accomplish some relatively minor things, usually only in a controlled laboratory environment. Or it would have to rely on external 'feedback' (which may be termed reinforcement, reward, punishment, etc.) in order to learn to select right actions, as in reinforcement learning in artificial intelligence (Sutton and Barto, 1998). But the requirement of external feedback begs the question of how such a signal is obtained in the natural world. In contrast, with the motivational system as an integral part of a cognitive architecture, it generates such feedback internally (from internal and external inputs) and learns on that basis (without requiring special external feedback, as will be discussed later). Furthermore, motivational processes are also important for social interaction. Each agent (human or robot) in a social situation carries with it its own needs, desires and motivations. Social interaction and cooperation are possible in part because agents can understand and appreciate each other's (innate or acquired) motivational structures (Snyder & Stukas, 1999; Tomasello, 1999). On that basis, agents may find ways to cooperate.

Similarly, without metacognitive control, an agent may be blindly single minded: it will not be able to flexibly and promptly adjust its own behaviour. The ability of an agent to 'reflect' on, and to modify dynamically, their own behaviour is important to cope effectively with complex environments. Social interaction is also made possible by the ability of agents to 'reflect' on, and to modify dynamically, their own behaviours (Tomasello, 1999). Metacognitive self-monitoring and regulation enable agents to interact with each other and with their environments more effectively, for example, by avoiding social impasses – impasses that are created because of the radically incompatible behaviours of multiple agents (see, e.g. Sun, 2001).

The main thesis of this article is that there are deeper layers of control of human behaviour and complex interactions amongst motivational, cognitive and metacognitive processes in the human mind, and these need to be taken into account in developing comprehensive models of the mind such as cognitive architectures. Note that here we limit our discussion to psychologically realistic models (which should be capable of capturing nuances of human psychology). In the remainder of this article, first some general background ideas are discussed, which lead to hypotheses concerning the existence and the structure of deeper control of behaviour. Then these ideas are solidified into a theoretical framework in the form of a cognitive architecture. Some evidence (including existing simulations) is presented (in an informal way) that shows how the framework helps to account for various issues and phenomena. Some further discussions and concluding remarks complete the paper.

2. Different views of behavioural control

In relation to the control of human behaviour, there are many questions that one may ask:

- Fundamentally, what determines what one chooses to do at each moment (beside external stimuli)?
- Are there multiple 'layers' of control of human behaviour? That is, beyond direct control in terms of moment-to-moment action selection, are there deeper, more fundamental 'layers' of control of human behaviour?
- What is the role of motivational processes in controlling behaviour?
- What is the role of metacognitive processes in controlling behaviour?

And many other related questions. It is important to address these issues in the development of psychologically realistic cognitive architectures, if such cognitive architectures are to become comprehensive theories of the mind in computational forms or all-encompassing frameworks for cognitive modelling.

In human everyday activities, behaviours are often generated directly and reactively, without involving elaborate processes (as argued by, e.g. Brooks, 1991; Heidegger, 1962). According to this view, everyday activities are largely made of habitual sequences of reactive behavioural responses or reactive routines (Sun, 2002). Correspondingly, in AI, some have advocated a reactive/situated/embodied approach towards building intelligent systems, which has been discussed at a theoretical level by, for example, Brooks (1991), Bickhard (1993), and Clancey (1997), among others.

On the other hand, there have been strong arguments for more complex cognitive processing and for deeper (motivational and metacognitive) control. Leaving aside the issue of the need for complex symbolic computation and representation (which have been discussed before; see, e.g. Fodor & Pylyshyn, 1988; Sun, 1994, 2002), we will look specifically into issues surrounding motivational and metacognitive control.

First let us examine the issue of (separate) motivational processes. In the work on animal behaviour, separate representation of motivation has been very much taken for granted. There have been ideas concerning motivation from ethology, such as those discussed by Toates (1986), McFarland (1989), and others. The notion of instinct and the notion of drive have been proposed. Hull (1951) developed a specific notion of drive as the primary representation for motivation, in which drives arose from need states, and behaviours were driven so as to eliminate need states.

From another perspective, in human psychology, Murray (1938) proposed a set of essential needs, from the physiological to the social (such as food or social belongingness). Reiss (2004) proposed a similar set. Maslow (1987) developed a hierarchy of needs: Lower-level needs were believed to have higher priority; when lower-level needs were satisfied, higher-level needs became prominent. This view of motivation required somewhat elaborate representation (a hierarchy). In other work on human motivation, elaborate constructs were also often specified (see, e.g. Weiner, 1992 for a variety of proposals). In all, these different proposals all point to a somewhat separate motivational process.

The need for motivational processes was also highlighted by Dayan (2001). He argued that existing reinforcement learning algorithms could not take into consideration physiological (and other) needs without additional learning, while natural organisms were capable of taking into account changes in needs without additional learning. A separate representation of motivation (needs) was necessary to capture the motivationally appropriate behavioural selection by natural organisms. In an empirical study, Berridge and Schulkin (1989) first gave rats sucrose and saline solutions with one of a bitter and a sour taste. They then induced a strong physiological need for salt (for the first time in the life of the rats). Given a choice, the rats preferred to drink the water with the flavour that had been previously associated with salt. The rats showed positive hedonic reactions towards the flavour associated with salt, whereas before pairing it was treated as being aversive. Such preferences were evident with no opportunity for learning. The key point was that behaviour was elicited only when it was motivationally appropriate, which suggested that behavioural selection was mediated by a motivational system (Grossberg, 1971).

In their treatment of human decision-making, Stout, Busemeyer, Lin, Grant, and Bonson (2004) included a separate representation of motivation. In their model, motivation underpinned decision-making by creating a valence, which led to decisions. Busemeyer, Townsend, and Stout (2002) developed a system of motivation, in which there was a desired level for each attribute, and motivational representations interacted with other factors in reaching a decision. Leven and Levine (1996) also showed how motivational representation mediated human decision-making. These models highlighted the need for a separate motivational system.

According to Ryan and Deci (2000), Kernis and Goldman (2005), and others, different types of motivations are present, ranging from engaging in an activity purely for the enjoyment, to engaging in an activity merely for obtaining a tangible reward or avoiding a punishment. Criticisms of popular models of human motivations (e.g. in economics) have focused on their overly narrow views regarding motivations, for example, solely in terms of reward and punishment (e.g. economic incentives and disincentives). Many critics opposed this overly narrow approach (Burns & Roszkowska, 2006; Pennisi, 2005). Complex social motivations, such as desire for reciprocation, seeking of social approval and interest in exploration, also shape human behaviour (Maslow, 1962; Murray, 1938; Sun, 2009). By neglecting these motivations, the understanding of some key social and behavioural issues may be hampered.

Similar criticisms apply to work on reinforcement learning in AI (Sutton and Barto, 1998). There have also been demonstrations of the advantages of incorporating motivational processes in terms of enhancing performance (e.g. survival), for example, by Tyrell (1993) and Scheutz and Sloman (2001). Again, these criticisms pointed to the need for complex (likely separate) motivational processes.

Turn now to metacognitive processes. According to Flavell (1976), metacognition refers to 'one's knowledge concerning one's own cognitive processes and products, or anything related to them'; metacognition also includes 'the active monitoring and consequent regulation and orchestration of these processes in relation to the cognitive objects or data on which they bear, usually in the service of some concrete goal or objective' (p. 232). Metacognition has been extensively studied in cognitive psychology (e.g. Mazzoni & Nelson, 1998; Metcalfe & Shimamura, 1994; Reder, 1996).

Metacognitive processes have often been conceived as specialised mechanisms that are separate and standalone for the purpose of monitoring and regulating cognitive processes. Such a conception has been explicitly stated in some theoretical treatments (e.g. Darling et al., 1998; Nelson & Narens, 1990), as well as implicitly assumed in many empirical studies (e.g. Metcalfe & Shimamura, 1994; Schneider, 1998).

It has been suggested that the prefrontal cortex has a lot to do with metacognitive functioning (e.g. Bensen, 1994). The major empirical evidence that metacognitive control might involve a separate (sub)system was the effect of a frontal lesion that showed that the control processes could be selectively impaired (Umilta & Stablum, 1998). Luria (1966) presented related evidence that the frontal lobe contained a (sub)system that programmed and regulated activities. See also Shallice (1988) and Stuss and Benson (1986).

In Koriat and Goldsmith's (1998) theory of metacognition, elaborate mechanisms of metacognition, separate from regular cognitive mechanisms, were hypothesised. In Shallice's (1988) theory, there was a rather separate 'supervisory attentional system', which dealt with some metacognitive functions such as schema selection (especially in non-routine situations). In Grafman's (1994) theory, there were a number of 'managerial knowledge units', which were separate from regular schemas and performed metacognitive control. In Atkinson's (1964) expectancy-value theory, elaborate representations of self, culture and experiences led to the representation of expectancies and values for the purpose of regulating behaviours, which was separate from regular representations. However, there have also been some criticisms of the assumption of a separate metacognitive controller.

In some computational models of metacognition-related phenomena, separate metacognitive constructs were also hypothesised. For example, PRODIGY incorporated problem solving, planning and multiple learning methods (Carbonell, Knoblock, & Minton, 1991). The problem solver searched over a problem space until a node was found that satisfied the top-level goal. Control (i.e. metacognitive) rules were used for improving search efficiency, improving solution quality and directing the problem solver along normally unexplored paths. CogAff (Sloman & Chrisley, 2005) combined reactive, deliberative and meta-management (i.e. metacognitive) functions. Separate representations were present for metacognitive purposes in CogAff. However, note that the term 'metacognition' has been used to denote different things and these different senses of metacognition may be carried out in different ways (more later).

So, should a computational cognitive architecture be based on a purely reactive (cf. Bickhard, 1993; Brooks, 1991) or a more complex (e.g. motivational or metacognitive) model of the mind? In the foregoing discussion, we looked into existing literatures broadly in search of some conclusions. Alternatively, we can also compare alternative ways of modelling and simulating empirical data to shed some light on psychological processes, which has been done in our previous work, such as Sun (2002, 2016), and will not be repeated here.

3. Structures of behavioural control

Now we proceed to address questions regarding the structure of deeper behavioural control involving motivational and metacognitive processes. For instance,

- Is the control (motivational or metacognitive) completely explicit or is it mixed in terms of involving both explicit and implicit processes? What are the respective roles of implicit and explicit processes in such control?
- Computationally speaking, what are the essential mechanisms and processes in such control?
- What are the essential representations involved in motivational processes?
- What are the essential representations involved in metacognitive processes?
- How do different types of processes (motivational, metacognitive and regular) interact and how can their interaction be captured computationally?
- What is executive control and how does it relate to metacognitive and motivational processes?

To address these issues, computational modelling (e.g. with a cognitive architecture) would be useful. In this section, we will attempt to analyse some of these issues, and then in the next section instantiate our analysis in the form of a computational cognitive architecture.

First, some general perspectives on the issue of implicit vs. explicit psychological processes are needed, which are important to the structure of behavioural control. Are psychological processes explicitly represented, or are they implicitly represented (in some ways)? A key assumption here is the dichotomy of the implicit and the explicit, which has been extensively argued for before (see Sun, Slusarz, & Terry, 2005; as well as Sun, 2002, 2016), Generally speaking, implicit processes are not directly consciously accessible and more 'holistic', while explicit processes are more accessible and more crisp (Reber, 1989; Sun, 2002). The dichotomy can be justified psychologically, by the voluminous empirical studies of implicit and explicit learning, implicit and explicit memory, implicit and explicit perception and so on (Cleeremans, Destrebecqz, & Boyer, 1998; Evans & Frankish, 2009; Reber, 1989; Seger, 1994; Sun, 2002). In social psychology, there are some related dual-process models (e.g. Chaiken & Trope, 1999). Denoting more or less the same distinction, these dichotomies serve as justifications for the more general notions of implicit vs. explicit cognition (which is the focus of the Clarion cognitive architecture, to be detailed later). See Sun (2002, 2016) and Sun et al. (2005) for more extensive treatments of this distinction.

Duality of this kind is present in both regular processes of cognition and 'deeper' layers of control, that is, motivational and metacognitive processes (Sun & Mathews, 2012; Woike, 1995). Computational modelling needs to capture details of the duality in all of these processes.

Now look into specifically the issue of explicit vs. implicit motivational processes. On one hand, it is hard to imagine that there is no explicit representation of goals, since all the evidence points to the contrary. For example, computational modelling by Anderson and Lebiere (1998), Laird (2012) and so on all pointed to the need for an explicit goal representation in a cognitive architecture. But, on the other hand, the inner workings of drives, needs, or desires are certainly not explicit, that is, not readily accessible (Hull, 1951; Sun, 2009; Toates, 1986; Weiner, 1992). So, it seems reasonable to assume that the idea of dual representation is appropriate in this context. Some combinations of explicit and implicit representations of goals and drives are needed.

We hypothesised (Sun, 2009, 2016) that the explicit motivational representation consists of explicit goals. They are necessary because explicit goals provide specific, tangible and articulated motivations for actions. Explicit goal setting also allows more behavioural flexibility and formation of expectancies (Epstein, 1982). Moreover, it may sometimes be necessary to compute a match of a state of the world to the goal, so as to discern the progress in achieving the goal and to possibly generate context-dependent reinforcement signals. This match may be facilitated using explicit representation of goals. Furthermore, implicit drive states change from moment to moment, but explicit goal representation is more persistent and longer lasting. In many circumstances, persistence in goal attainment is needed. In addition, explicit goal representation enables explicit cognitive processes to work on these goals and their attainment (in addition to having implicit processes directing behaviour).

Implicit motivational processes are primary, and more essential than explicit processes involving explicit goals. Human motivational processes are known to be highly complex and varied (Weiner, 1992), and apparently cannot be captured with only explicit goal representation (e.g. as in Anderson & Lebiere, 1998). For example, the interaction of drives, especially their combinations, requires more complex and less explicit processes (McFarland, 1989; Tyrell, 1993). Their changes over time, which are often gradual and dynamic, also require more quantitative and graded representation. Given that, it is natural to hypothesise that implicit motivational processes are fundamental; on that basis, explicit goal representation arises, which clarifies and gives some directions to implicit motivational and behavioural dynamics. Castelfranchi (2001), for example, discussed such implicit-to-explicit motivational processes (Sun, Merrill, & Peterson, 2001).

Empirical evidence from social psychology also indicated the duality of human motivation. For example, Woike (1995) showed how implicit and explicit motives might have different effects on memory recall. Wood and Quinn (2005) explored the duality of motivation in everyday life, and the relationship between implicit and explicit motivations, in ways analogous to the analysis of implicit and explicit cognitive processes in Sun et al. (2005). Strack and Deutsch (2005) expressed a similar view, describing what we have termed top-down and bottom-up influences (implicit motivation affecting explicit motivation and vise versa; Sun et al., 2005). Norton, Vandello, and Darley (2004) showed that people might be motivated implicitly by questionable criteria but then masked their implicit biases through engaging in casuistry explicitly. Hing, Chung-Yan, Grunfeld, Robichaud, and Zanna (2005) also demonstrated how implicit and explicit motivations might diverge and consequently how they might counter-balance each other. Adams, Wright, and Lohr (1996) even found that an individual's implicit and explicit motivations could be diametrically opposed.

Summarising the discussion so far, motivational processes are neither necessarily explicit, nor necessarily implicit. They are likely the results of the interaction between implicit and explicit processes, the same as regular cognitive processes (as has been argued by Sun, 2002, 2016; Sun et al., 2005).

Turn now to metacognitive processes. These processes have traditionally been portrayed as explicit processes that involve deliberative reasoning (e.g. Darling et al., 1998; Nelson & Narens, 1990). However, there have been various experimental indications and theoretical arguments that metacognition may be in part implicit. For example, Reder and Schunn (1996) interpreted the results from experiments in an arithmetic domain as indicating that feeling-of-knowing (FOK) judgements were implicit. This was because results indicated that feeling of knowing was an error-prone associative process, which did not involve detailed analysis of terms involved (e.g. exposure to '23 * 14' led to a higher FOK for '23 + 14').

Likewise, Metcalfe (1986) demonstrated the failure of 'warmth' judgement (feeling of closeness to solutions) in predicting how close a participant was to solving a problem: participants who successfully solved a problem had almost constant warmth ratings, while participants who failed to solve a problem might have rising warmth ratings. Further evidence of a dissociation between metacognitive judgements and actual performance was summarised by Schneider (1998), Reder (1996) and others. Schneider (1998) pointed out that feeling-of-knowing judgements varied with methods of assessment and with stimuli materials. It was also found that predictive accuracy of ease-of-learning (EOL) judgements varied with tasks. A general finding was that such judgements were not perfect predictors. The above facts point to the possibility that metacognition is largely implicit.

In social psychology, Bargh (1997) and Wegner (1994) showed that much of the regulatory monitoring processes went on outside of awareness. For instance, Bargh (1997) described considerable evidence that individuals could accomplish goals and monitor their progress without realising they were doing so. Wegner (1994) described an unconscious monitoring process for unwanted thoughts, which interacted with a conscious control process. Again, there is the separation of implicit and explicit metacognitive processes.

Beyond implicit metacognition, there are explicit elements in metacognition. For example, deliberative metacognitive reasoning does occur (Gentner & Collins, 1981) with the involvement of explicit processes. Explicit selection of strategies in reading comprehension is also a known phenomenon (Forrest-Pressley and Waller, 1984). In social psychology, there have been similar findings.

Reder (1987) posited that metacognitive judgement invoked implicit (similarity-based) processes first and then a more explicit, deliberative and analytical process. Narens, Graf, and Nelson (1996) also showed that some metacognitive judgements were equally predictive of explicit and implicit memory, and thus metacognitive judgements might be the result of both explicit and implicit processes.

Therefore, like motivational processes, metacognitive processes are neither necessarily explicit, nor necessarily implicit. It is likely a combination of implicit and explicit processes, the same as regular cognitive processes (Sun, 2002, 2016).

We will now look into a number of other issues related to motivation and metacognition. Motivations are determined (and changed) on the basis of environmental states and internal states. For example, Dethier (1966) and Epstein (1982) discussed the importance of both internal states and external stimuli in determining and changing motivational states. Furthermore, McFarland (1989) proposed a model of motivational force based on a product of external and internal factors (i.e. a product of deficit and cue strength). Tyrell (1993) further developed this model. Aarts and Hassin (2005) discussed mechanisms for the unconscious process of motivation, dependent on both internal and external factors. They showed that people could automatically (unconsciously) infer other people's motivation and then adopt and pursue them.

Hertel, Kerr, and Messe (2000) showed how being in a team situation affected motivations. Their work illustrated how different social environments led to different social motivations. Tauer and Harackiewicz (2004) similarly showed how cooperation and competition enhanced motivation. Iyengar and Lepper (1999) showed how cultural environments were internalised as intrinsic motivations of participants, which in turn determined specific motivations in specific situations.

However, motivations may also be elicited from internal sources (in addition to external factors). For example, they can be elicited through explicit reasoning (e.g. concerning a social situation). It is known as cognitive appraisal in the emotion research literature (e.g. Hudlicka & Fellous, 1996; Sun, Wilson, & Lynch, 2016).

In these accounts, both internal and external factors were the determinants of motivation. Therefore, the idea of motivational changes based on environmental and internal states needs to be taken into serious consideration in developing a cognitive architecture.

Going further, based on motivational states, metacognitive regulation may be carried out. Looking into the literature, there is indeed evidence in support of metacognitive regulation based on motivational states (as well as on other factors such as performance monitoring and current environmental states).

Neuberg, Kenrick, Manor, and Shaller (2005) discussed the link from motivation to selective attention. They showed how fundamental social motives influenced early stage perceptual and cognitive processing. Maner et al. (2005) showed that different motivations led people to perceive different features in faces of other people. Strack and Deutsch (2005) showed similar findings.

Kanfer and Ackerman (1989) investigated the effect of goal setting on training effectiveness and the interaction among goal setting, training content and ability. Their results demonstrated that motivation had a lot to do with skill learning: Motivations led to the setting of cognitive parameters that constrained performance.

There have also been relevant findings in social psychology. Maheswaran and Chaiken (1991) showed how motivations determined processing modes – whether systematic processing (likely explicit) or heuristic processing (likely implicit) was adopted. Chen, Shechter, and Chaiken (1996) also showed how motivations affected processing modes in a social interaction setting: Impression-motivated participants adopted a 'go along to get along' strategy, while accuracy-motivated participants adopted a more systematic processing mode. In all of these cases, motivations lead to selecting corresponding modes of cognitive processing.

Note that metacognitive regulation and control are often heavily dependent on on-line monitoring: the current performance and its relation to the desired level of performance (Narens et al., 1996). Given all of the above evidence, metacognitive regulation based on motivational states (as well as other factors) is well supported.

Finally, routine/subroutine structures induced by goals should be examined. Sequentiality is an essential behavioural characteristic (Lashley, 1951), which needs to be captured, and routine/subroutine structures induced by goals are useful for achieving sequentiality (Sun, 2002).

A number of mechanisms that provide mechanistic underpinnings for a variety of animal behaviours have been proposed. Fixed action patterns (FAPs), innate releasing mechanisms (IRMs) and modal action patterns (MAPs) are all instances of such constructs. For example, FAPs are self-contained entities in that each has a source of motivational energy that activates a specific sequence of behaviours (a routine/subroutine), under some particular stimulus conditions. According to Lorenz (1950) and Tinbergen (1951), each species was equipped with a sufficient variety of FAPs to ensure an appropriate response in any normal circumstances.

Roughly related to such animal motivational constructs, the use of a goal stack in human cognition was proposed in early cognitive architectures (Anderson & Lebiere, 1998). A goal stack allows for routines/subroutines (Sun, 2002). Once a goal is pushed onto the stack, a routine/subroutine for accomplishing the goal is automatically initiated (Lorenz, 1950), through selecting actions suitable for accomplishing the goal. An initiated routine can keep running, until interrupted or terminated. During its running, a higher priority goal may be pushed onto the goal stack. The current routine can then be suspended, and the routine for the new goal be carried out. At the termination of the new routine, the previous routine may be resumed. A goal may spawn subgoals, by pushing these subgoals onto the stack.

There are alternatives beyond such a simple approach. For example, in robotics, there have been various proposals concerning 'layered architectures' (e.g. Gat, 1998). These architectures in general divide action control into multiple layers: for instance, (1) the controller, which takes actions reactively in response to environmental input in accordance with some behavioural routines, (2) the sequencer, which selects among different behavioural routines to be carried out by the controller, and (3) the deliberator, which plans out future courses of actions and directs the sequencer to act accordingly. The division of labour among different components is useful for maintaining both fast responses and behavioural flexibility.

More sophisticated goal structures and (sub)routine mechanisms are needed in cognitive architectures. For instance, in a more sophisticated cognitive architecture, goals may emerge from competitions among different needs and desires, goals may change in various ways (including in a stack-like fashion as well as other possibilities) and so on. Routines may have both of the following two properties: persistence and interruptibility (Tyrell, 1993); the interplay of these properties may need to be addressed (Sun, 2009, 2016). In addition, the initiation of routines (e.g. by setting goals), the routines themselves, and the termination of routines can all be learned (in addition to being hand-coded, as in many existing cognitive architectures). Their learning may be done through autonomous trial-and-error exploration, instructions, imitations and other means (e.g. Sun, 2016; Sun & Sessions, 2000).

Metacognitive regulation affects routine behaviour, for example, by switching from one routine to another (Norman & Shallice, 1986). Metacognition also helps to form routines as well as to adapt routines

(Kanfer and Ackerman, 1989). In these processes, routine behaviours are modulated in accordance with motivational states (as well as other factors); they serve activated motivations (or they may even become motives themselves; Herrstein, 1977; Toates, 1986).

4. A theoretical framework

On the basis of the foregoing discussions, we are ready to describe a theoretical framework in the form of a computational cognitive architecture that includes motivational and metacognitive control of cognitive processes that lead to behaviour. To foster integration and avoid fragmentation into narrow and isolated sub-disciplines, it is necessary to consider the overall architecture of the mind that incorporates, rather than excludes, important elements such as motivations and metacognition. Furthermore, it is clearly beneficial to translate into computational and architectural terms the understanding that has been achieved of the interactions among cognitive, metacognitive and motivational aspects of the mind (Dai & Sternberg, 2004; Sun, 2009), and generate further hypotheses in the process. In so doing, we will be able to produce a more complete picture of the structure of the mind.

In developing a cognitive architecture, details concerning motivation and metacognition need to be addressed. Motivational mechanisms are concerned with why an agent does what it does – on what basis an agent chooses the actions that it takes. Simply saying that an agent chooses actions to maximise gains, rewards, reinforcement or payoffs leaves open the question of what determines gains, rewards, reinforcement or payoffs (Sun, 2016; Toates, 1986; Weiner, 1992). Within a cognitive architecture, the relevance of motivational processes lies in the fact that they provide the context in which performance and learning are carried out.

Closely tied to motivational mechanisms, metacognitive mechanisms are also needed in a cognitive architecture (Carver & Scheier, 1990; Mazzoni & Nelson, 1998). Metacognitive mechanisms are for monitoring, controlling and regulating cognitive processes, for the sake of performance, in the form of setting specific goals, setting essential parameters, interrupting and changing on-going processes and so on.

Our theoretical framework has been fully expressed in a computational cognitive architecture, namely Clarion (Sun, 2002, 2016). Clarion is intended for capturing all the essential psychological processes within an individual. Below, the framework will be sketched. Then, some limited details of the cognitive architecture will be described.

4.1. Overview

Clarion consists of a number of subsystems, with a dual representational structure in each subsystem (implicit vs. explicit representation). Its subsystems include the action-centred subsystem (the ACS), the non-action-centred subsystem (the NACS), the motivational subsystem (the MS), and the metacognitive subsystem (the MCS). The role of the action-centred subsystem is to control actions, regardless of whether the actions are for external physical movements or for internal mental operations. The role of the non-action-centred subsystem is to maintain and utilise general knowledge, either implicit or explicit. The role of the motivational subsystem is to provide underlying motivations for action, 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.[1]

Each of these interacting subsystems consists of two levels of representation (i.e. a dual representational structure): Generally, in each subsystem, the top level encodes explicit knowledge and the bottom level encodes implicit knowledge. This distinction has been argued for before (Cleeremans et al., 1998; Reber, 1989; Seger, 1994; Sun, 2002). The relatively inaccessible nature of implicit knowledge may be captured by distributed representation provided, for example, by a backpropagation network (Rumelhart, McClelland, & The PDP Research Group, 1986). Distributed representational units (in the hidden layers of a backpropagation network) are capable of accomplishing computation but are generally not individually meaningful (Rumelhart et al., 1986; Sun, 1994). In contrast, explicit knowledge may be captured in computational modelling by localist-symbolic representation (Clark & Karmiloff-Smith, 1993), in which each unit is more easily interpretable and has a clearer conceptual meaning. The dichotomous difference between the two different types of knowledge leads naturally to a two-level (dual-representational) architecture (Sun, 1994, 2002), whereby each level uses one kind of representation and captures one corresponding type of process (implicit or explicit).

Figure 1 presents a sketch of this basic architecture, which includes the four major subsystems interacting with each other (Sun, 2002, 2016). The following four subsections will describe, one by one and in more detail, these four subsystems.

4.2. The action-centred subsystem

The action-centred subsystem (the ACS) captures the action selection of an individual when interacting with the world. This subsystem is the central part of Clarion.

Within the subsystem, the process for action selection is essentially as follows: Observing the current (observable) input state of the world, the two levels within the subsystem (implicit or explicit) make their separate decisions in accordance with their respective procedural knowledge and their outcomes are somehow 'integrated'. Thus, a final (stochastic) selection of an action is made and the action is then performed. The action changes the world in some way. Comparing the changed input state with the previous input state somehow, the person learns. The cycle then repeats itself.

While the bottom level is implemented with neural networks involving distributed representation (Rumelhart et al., 1986), the top level is implemented with rules using localist representation. At the top level, 'chunk' nodes are used for denoting concepts (a localist representation). A chunk node connects to its corresponding microfeatures (dimensional values) represented as separate nodes in the bottom level (distributed representation). According to Kant, the role of a schema is to provide an image for a concept. In this case, distributed microfeature-based representation of a concept at the bottom level

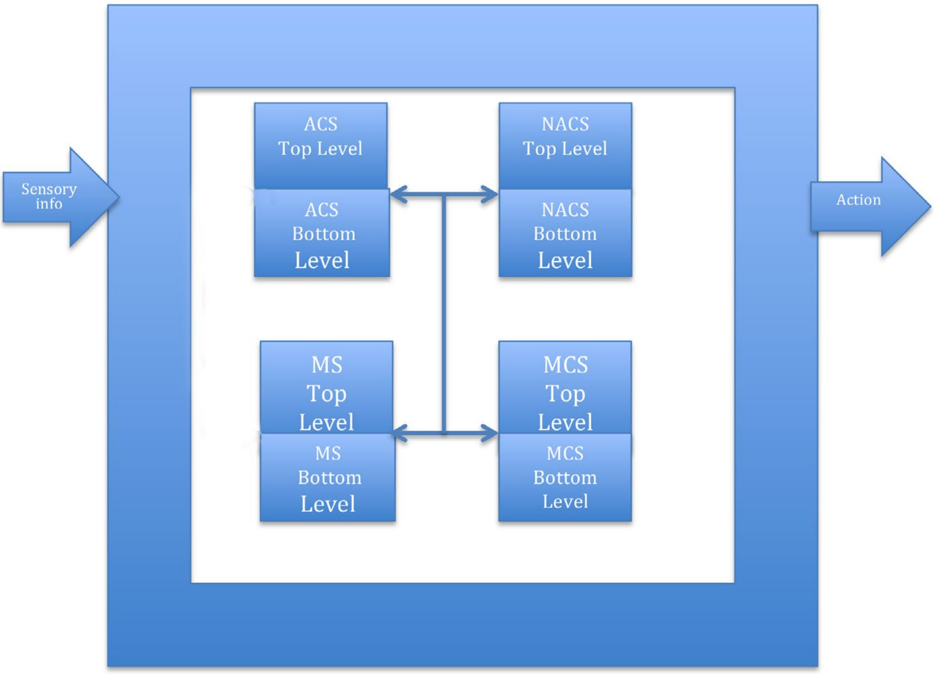

Figure 1. The Clarion cognitive architecture.
Notes: ACS denotes the action-centred subsystem, NACS the non-action-centred subsystem, MS the motivational subsystem and MCS the metacognitive subsystem.

indeed provides a mental 'image' for the concept that the corresponding chunk node at the top level represents (Ricoeur, 1981; Sun, 1994, 2002, 2016).

The input consists of several sets of information: sensory information (environmental or internal), the current goal and so on. The sensory input and the current goal are both important in deciding on an action. The input to the bottom level is represented as a set of microfeatures (constituting distributed representation). The output is the action choice.

At the bottom level, in a neural network encoding implicit procedural knowledge, actions are selected based on their *Q* values. A *Q* value is an evaluation of the 'quality' of an action in a given input state. At each step, given the input state, the *Q* values of all the actions 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 (Luce's choice axiom; Watkins, 1989).

For learning implicit procedural knowledge at the bottom level (as represented by the *Q* values), a reinforcement learning algorithm may be used. The neural network computing *Q* values is gradually tuned, on-line, through successive updating, which enables reactive sequential behaviour to emerge through trial-and-error interaction with the world. In this learning setting, there is no need for a priori knowledge or external teachers providing desired input/output mappings (except a need for a feedback or reinforcement signal, which will be addressed later). Such implicit learning may be justified cognitively. For instance, Cleeremans (1997) argued that implicit learning could not be captured by symbolic models but neural networks. Sun (2002) and Sun et al. (2005) made similar arguments.

Explicit knowledge at the top level can also be learned in a variety of ways (with rules using localist representation). One-shot learning (e.g. based on hypothesis testing) during interaction with the world is appropriate (Bruner, Goodnow, & Austin, 1956; Sun et al., 2001). For instance, a 'bottom-up' learning process (Karmiloff-Smith, 1986; Sun et al., 2001) may take place, which utilises information from the bottom level. 'Top-down' learning can also occur to assimilate the explicit knowledge of the top level into the bottom level.

For stochastic selection of the outcomes of the two levels, at each step, with a certain probability, the outcome of the bottom level is used. Otherwise, the outcome from the rules of the top level is used. There exists some psychological evidence for such intermittent use of rules (Sun, 2002; Sun et al., 2005).

4.3. The non-action-centred subsystem

The non-action-centred subsystem (the NACS) deals with declarative knowledge, which is not action-centred, for the purpose of making inferences about the world. 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' at the bottom level. Its operation is under the direction of the action-centred subsystem. It captures traditional notions of beliefs/knowledge and reasoning (Sun, 2016).

On one hand, at the bottom level of this subsystem, 'associative memory' networks encode implicit declarative (non-action-centred) knowledge (Rumelhart et al., 1986). On the other hand, at the top level, explicit declarative (non-action-centred) knowledge is stored. As in the action-centred subsystem, 'chunk' nodes (denoting concepts) at the top level are linked to microfeatures represented at the bottom level. Additionally, at the top level, links between chunk nodes encode explicit associative rules.

As in the action-centred subsystem, various kinds of learning take place. For example, top-down or bottom-up learning may take place, either to extract explicit knowledge for the top level from the implicit knowledge in the bottom level, or to assimilate the explicit knowledge of the top level into the bottom level.

4.4. The motivational subsystem

The motivational subsystem (the MS) of Clarion is concerned with why an individual does what he/she does (Toates, 1986; Weiner, 1992). The relevance of this subsystem to the action-centred subsystem lies primarily in the fact that it provides the context in which the goal and the reinforcement are determined. It thereby influences the working of the action-centred subsystem (and by extension, the working of the NACS).

Hull (1951) developed a notion of 'drives' – an implicit, pre-conceptual representation of motives. In his view, drives arose from need states, behaviours were driven so as to eliminate need states, and drive reduction was the basis of reinforcement. Although Hull's conception of drive had significant explanatory power, it failed to capture many motivational phenomena. A more general notion is therefore needed. In Clarion, a generalised notion of 'drive', different from the strict interpretation of drives (e.g. as physiological deficits that require to be reduced by corresponding behaviour), is adopted. In our sense, drives denote internal needs or desires of all kinds that likely may lead to corresponding behaviours, regardless of whether the needs are physiological or not, whether the needs may be reduced by the behaviours or not, or whether the needs are for end states or for processes. It is a generalised notion that transcends controversies surrounding the stricter notions. This notion is adopted, because we need to account for (1) context-dependent and (2) persistent but interruptible focus of behaviour, as well as other properties mentioned early on (Sun, 2009).

As argued before, the idea of dual representation applies to the motivational representations as well as to other mental representations (Forgas, Williams, & Laham, 2005; Schooler & Schreiber, 2005; Woike, 1995). Therefore, in this subsystem, explicit goals (such as 'find food'), which are essential to the working of the action-centred subsystem, may be generated based on implicit drives (e.g. 'being hungry'). Explicit goals derive from, and hinge upon, implicit drives. See Figure 2 for a sketch of the subsystem.

In particular, within the subsystem, 'primary drives' refer to those drives that are essential to an individual and are most likely built-in (hard-wired) to a significant extent to begin with. Some sample low-level primary drives include: *food, water, reproduction* and so on (McDougall, 1908; Murray, 1938). Beyond such low-level primary drives (concerning physiological needs), there are also high-level primary drives: for example, *dominance and power, fairness* and so on.

The primary drives (low-level and high-level together) may be roughly explained as in Table 1. This set of primary drives has been extensively explored and justified in prior writings (Sun, 2009). Note that there have been a variety of such drives being proposed by various researchers. For example, the work of McDougall (1908), Lewin (1936), Murray (1938), Maslow (1987), Hull (1951), McClelland (1951), and Reiss (2004) has been well known. This set of primary drives is essentially the same as Murray's (1938), with only a few differences. Likewise, this set of drives is similar to Reiss's (2004) set, but with some differences. So, the prior work by these researchers in justifying their frameworks may be applied, to a significant extent, to this set of drives as well (Sun, 2009).

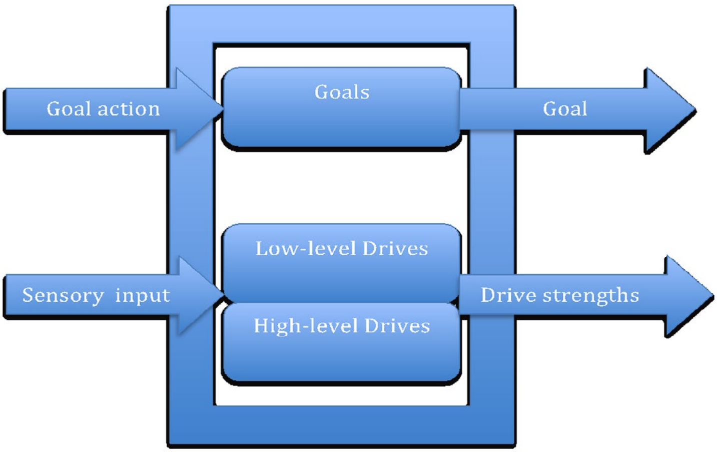

Figure 2. The structure of the motivational subsystem.

Table 1. Primary drives in Clarion.

Food	The drive to consume nourishment
Water	The drive to consume liquid
Sleep	The drive to rest and/or sleep
Reproduction	The drive to mate
Avoiding danger	The drive to avoid situations that have the potential to be or already are 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 or groups
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 and serve a person of a higher status of some kind
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
Honour	The drive to follow social norms and codes of behaviour 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 organise, or to structure (e.g. one's environment)
Curiosity	The drive to explore, to discover and to gain new knowledge

Besides these primary drives, which are built-in and relatively unalterable, there are also 'derived' drives, which are secondary, changeable and acquired mostly in the process of satisfying primary drives.

These drives may be activated to different extents in each moment. The strengths of drives in a given setting are calculated using a neural network based roughly on the product of the input stimulus and the internal 'deficit' (which represents an individual's intrinsic sensitivity and inclination toward activating a drive). The justifications for this may be found in a variety of literatures (Sun, 2009). See Sun (2016) for full details.

Note that the degree to which each of these drives is significant is variable. It may be modulated, to some extent, by individual psychological differences, individual physical/physiological conditions, cultural differences and so on. Nevertheless, the existence of these drives and their importance to the functioning of humans and human societies are believed to be universal, as has been discussed in the literature (including by clinical psychologists and psychoanalysts; e.g. Freud, 1915).

4.5. *The metacognitive subsystem*

Metacognition refers to one's knowledge (implicit or explicit) concerning one's own cognitive processes. Metacognition includes active monitoring and regulation of these processes (Flavell, 1976). In Clarion, the metacognitive subsystem (the MCS) is closely tied to the motivational subsystem and it monitors, controls and regulates cognitive processes. Control and regulation may be in the form of

- setting goals (which are then used by the ACS) on the basis of drives (Eccles & Wigfield, 2002; Latham & Pinder, 2005; Tolman, 1932),
- setting reinforcements (for learning within the ACS), on the basis of drives and goals (i.e. how drives and goals are satisfied; Dayan, 2001; Hull, 1951),
- information filtering (Derryberry & Tucker, 1994; Logan & Gordon, 2001; Neuberg et al., 2005),
- setting essential parameters of the ACS and the NACS (Kanfer and Ackerman, 1989),

and so on. It also includes buffers for keeping track of certain internal information (such as 'warmth' ratings; see Sun, Zhang, & Mathews, 2006), based on motivation and other factors. Structurally, this subsystem may be divided into a number of modules, as shown in Figure 3.

For instance, within the goal module, in order to select a new goal, it first determines goal strengths for some or all of the goals, based on information from the motivational subsystem and the current sensory inputs. Then, a new goal is stochastically selected on the basis of the goal strengths. For the general notion of, and arguments in support of, goal setting on the basis of implicit motives (i.e. drives), see, for example, Tolman (1932), Deci and Staub (1980).

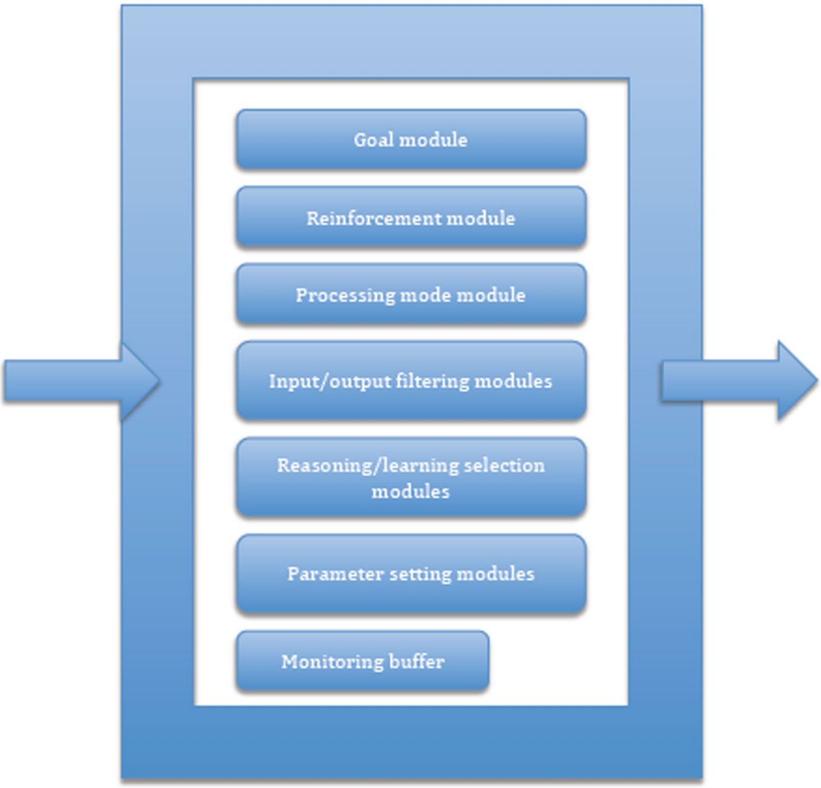
Goal module
Reinforcement module
Processing mode module
Input/output filtering modules
Reasoning/learning selection modules
Parameter setting modules
Monitoring buffer

Figure 3. The metacognitive subsystem.

For another instance, motivations may influence perception. Maner et al. (2005) showed that different motivations led people to perceive different features in faces of other people. Strack and Deutsch (2005) also described similar findings. The input filtering module accomplishes that.

This subsystem is comprised of two levels of implicit and explicit processes (the same as the other subsystems). In this subsystem, mostly, the bottom level takes effective control, and the top level has less impact on outcomes. This is because metacognitive control is usually fast and effortless (Reder & Schunn, 1996), and thus it is mostly implicit. However, under some circumstances, the top level can also exert strong influence (Forrest-Pressley and Waller, 1984). Furthermore, different modules of the subsystem may have different degrees of reliance on the bottom level.

4.6. Various interactions and deeper control

Clarion tackles issues not adequately addressed by other existent cognitive architectures, such as the implicit–explicit interaction, the cognitive–metacognitive interaction and the cognitive–motivational interaction. Within Clarion, there is the following closed loop:

> environmental input → motivation in the motivational subsystem → metacognition in the metacognitive subsystem → actions by the action-centered subsystem (possibly consulting the non-action-centered subsystem) → environmental changes

This loop mediates the complex interaction between an agent and its external environments, which cannot conceivably be captured by a simpler model, including the reactive (situated) view and the traditional symbolic cognitivist view. Both of these two views tended to ignore the important roles and details of motivational and metacognitive control.

Moreover, there are two intersecting arcs in this loop. Instead of going directly from stimuli to actions, one arc goes from stimuli (environmental input) to motivation (first to drives and then to goals), and then further to metacognitive control and regulation (which in turn affects actions); the other arc goes from stimuli and goals (as determined by the MS and the MCS) to actions. That is, for the sake of flexibility, there is a division of labour. This division of labour enables the mediation of direct stimulus-response relations by motivational and metacognitive processes (Dayan, 2001; Hull, 1951; Sun, 2009). It also enables the formation of routines in behaviour, by interjecting motivational structures between stimuli and actions (see the discussion of routines earlier).

The details of the Clarion cognitive architecture, including its computational details, have been extensively discussed elsewhere. Further details will not be provided here, in order to maintain the non-technical nature of the present article. See Sun (2002, 2016) for further information. Note also the limited modularity within Clarion: These 'modules' interact heavily with each other and furthermore, processes within different modules may be similar in many ways.

5. Roles of behavioural control

In philosophy of mind, instrumentalism is the view that propositional attitudes are not for 'scientific' investigations of the mind and brain, but assuming that an agent does have beliefs (as well as desires, goals and so on) is often a useful strategy in elucidating the inner working of the agent. Let us see what benefit the notion of deeper control (motivational and/or metacognitive) may have in shedding light on behavioural control in humans (and artificial agents), leaving aside for the time being 'scientific' questions such as the neurobiological underpinning of such control.

Clarion, as a framework, has been successful in simulating a variety of psychological tasks (see, e.g. Sun, 2002, 2016; Sun et al., 2001, 2005). It should be capable of capturing a wide range of motivational and metacognitive control phenomena. Simulations involving motivational and metacognitive processes have indeed been carried out. The inclusion of motivational and metacognitive constructs in Clarion also facilitates modelling and simulation in a practical way: Built-in motivational and metacognitive mechanisms speed up simulations involving motivation and metacognition by making the job of setting up simulations easier. Below we examine briefly a few examples.

5.1. Reactivity and motivation

As discussed by Heidegger (1962), Brooks (1991), Bickhard (1993), Clancey (1997), and others, in human everyday activities, behavioural responses are mostly generated without involving elaborate computation; that is, they are mostly generated reactively. Everyday activities are largely made of habitual sequences of reactive behavioural responses (i.e. reactive routines; Sun, 2002).

Reactivity of human behaviour entails relatively fixed responses, so that an individual does not have to re-compute responses every time a response is needed. Such reactivity is also direct, and does not involve elaborate mediating representations (Bickhard, 1993). Thus, they are mostly implicit (Sun, 2002). This view is consistent with voluminous research on implicit learning, implicit memory and so on (Evans & Frankish, 2009; Reber, 1989; Sun, 2002).

On top of implicit processes for reactive behaviour, there are of course explicit conceptual representations and explicit conceptual processes. But they are at a higher 'level'; that is, they constitute a separate component of the mind that likely functions on top of implicit processes, as stipulated in Clarion (as elaborated in Sun, 2002, 2016).

In Clarion, reactive accounts of behaviour control (e.g. Bickhard, 1993; Brooks, 1991) co-exist with motivational accounts (e.g. Toates, 1986; Weiner, 1992). At its core, Clarion may be purely reactive, without the necessity of relying on motivational constructs, as in the action-centred subsystem, which can run by itself (simply assuming, e.g. that the inputs from the MS and the MCS are constant). On the other hand, Clarion can involve complex motivational dynamics, as occurring in the motivational subsystem, including a bipartite representation of goals and drives. Thus, in a way, it synthesises the two world views.

Contrary to some existing theories, in our view, reactivity does not necessarily exclude motivational and metacognitive processes. To react, one may base one's reactions on one's own needs and desires. Reactions that are consistent with one's needs and desires are proper reactions (Tolman, 1932). Thus, needs and desires may lie at the very foundation of reactivity (as well as at that of conceptual representation and reasoning).

However, in some sense, reactivity does exclude a significant role for overly elaborate explicit processes in motivation, metacognition and so on. This is because of the speed of reactivity: Elaborate conceptual reasoning and the like are simply too costly in terms of processing resources and processing time. So, on one hand, there are more reactive, more implicit motivational and metacognitive processes; on the other hand, there are more deliberative, more explicit ones (Sun, 2009; Sun & Mathews, 2012). In routine reactions, generally speaking, implicit processes dominate. Implicit processes likewise dominate motivation and metacognition in such situations. In contrast, in deliberative situations, generally speaking, explicit processes dominate. In such situations, explicit processes may likewise dominate motivation and metacognition. Thus, reactivity constrains motivational and metacognitive processes (when in reactive modes), rather than excluding them.

For instance, with regard to S-R conditioning, Tolman (1932) pointed out the importance of motivation: "Stimuli do not, as such, call out responses willy nilly. … Rather learning consists in the organisms' 'discovering' or 'refining' what all the respective alternative responses lead to. And then, if, under the appetite-aversion conditions of the moment, the consequences of one of these alternatives is more demanded than the others … then the organism will tend, after such learning, to select and to perform the response leading to the more 'demanded-for' consequences. But, if there be no such difference in demands there will be no such selection and performance of one response, even though there has been learning" (p. 364). See also Grossberg (1971) regarding this point. Motivation, which Clarion captures, is important even for S–R conditioning.

We may contrast Clarion with a typical reactive model such as the subsumption architecture (Brooks, 1991). Unlike the subsumption architecture, in Clarion, there is no fixed subsumption hierarchy. Which response routine should dominate in a given situation is dependent on a host of internal and external factors in Clarion. Thus, Clarion offers more flexibility. Such flexibility is needed in order to capture observed behavioural flexibility of real organisms (especially humans). For a simple example, whether eating is preferred over drinking is dependent on the internal state (e.g. food deficit vs. water deficit) of an organism. The subsumption architecture is not set-up to deal with this kind and more complex kinds of flexibility.

5.2. Executive control and metacognition

Although empirical work explores 'executive control', it is sometimes unclear what is exactly meant by executive control and how it works (Logan, 2003). Detailed models are needed to substantiate it (Logan & Gordon, 2001). This notion should also be related to metacognitive and motivational issues.

In Clarion, 'executive control' is somewhat generalised and linked to some other functions; it also becomes more distributed. For example, the action-centred subsystem controls not only external actions but also some other cognitive processes (e.g. the NACS). That is, in Clarion, the control of internal processes and the control of external processes are one and the same, both resulting from the decisions of the action-centred subsystem. Its decision may be concerned with 'executive control' of memory and reasoning (within the NACS). On top of that, there is metacognitive monitoring and regulation through the metacognitive subsystem, which may alter the functioning of other subsystems. Some may regard some of these functions as also belonging to 'executive control'. In that sense, executive control is distributed between the action-centred subsystem and the metacognitive subsystem in Clarion.

Likewise, in the literature, the term 'metacognition' has been used to denote a number of different things (e.g. Mazzoni & Nelson, 1998; Metcalfe & Shimamura, 1994; Nelson & Narens, 1990). These different senses of metacognition may be carried out in different ways and in different subsystems within Clarion. The metacognitive subsystem of Clarion, as its name suggests, carries out many functions that have been considered to be part of metacognition (including setting up goals and internal reinforcements, as well as adjusting a number of major parameters; Sun, 2016). However, some of what has been referred to as 'metacognition' occurs in other subsystems of Clarion (in particular, the ACS and the NACS).

Specifically, to decide which activities to focus on, it is possible that, in the action-centered subsystem, in addition to goal setting, different modules responsible for different activities compete with each other and the winning module takes on its corresponding activities (Norman & Shallice, 1986). To decide which aspects of a task to focus on, in addition to filtering by the metacognitive subsystem (as described before), it is likely that decisions are made implicitly by the networks at the bottom level, while top-level explicit knowledge may have impact as well (Derryberry & Tucker, 1994; Neuberg et al., 2005). To decide which strategy to adopt in the face of a particular task (Reder & Schunn, 1996), either inter-module competition or intra-module implicit decision-making may be responsible at the bottom level, and explicit processes at the top level can affect strategy choices when explicit reasoning is used to select strategies (e.g. in relation to current goals). In addition, important strategic decisions may be made by specialised modules; after a strategy is selected by this module, it may be carried out by another module (i.e. hierarchical decision-making). Therefore, the process for strategic decisions varies, depending on the type, significance and granularity of a particular strategic decision.

In the same vein, some metacognitive knowledge evoked in experimental work may conceivably be found in the action-centred or the non-action-centred subsystem. For example, the 'feeling of knowing' judgement (Reder & Schunn, 1996) may be assessed in the non-action-centred subsystem through similarity-based reasoning processes (Sun, 2016). Some other types of metacognitive knowledge, such as the 'warmth' judgement (Metcalfe, 1986), may be registered in the monitoring buffer of the metacognitive subsystem (Sun, 2016; Sun et al., 2006). Metacognitive control, on the basis of such information, can be carried out either by the metacognitive subsystem or by the action-centred subsystem (especially when decisions to be made are at a relatively low level). A number of metacognitive simulations of specific experimental data have been carried out within Clarion; see, for example, Sun et al. (2006) and Wilson, Sun, and Mathews (2009).

5.3. Personality

Because a computational cognitive architecture should include all essential psychological mechanisms and processes, the interaction within it should be able to generate psychological phenomena of all kinds, which of course include personality-related phenomena (Sun & Wilson, 2014). Therefore, personality should be computationally explainable by a computational cognitive architecture. Motivational, metacognitive, action selection, reasoning and other processes capture the interaction of internal needs and external environmental factors in determining goals and actions by individuals. They capture the relative invariance within an individual in terms of behavioural propensities and inclinations at different times and with regard to different situations (social or physical), as well as behavioural variability (Sun & Wilson, 2014).

Among these processes, motivation is especially pertinent to personality (Read et al., 2010; Sun & Wilson, 2014). As discussed earlier, within the motivational subsystem, there is a set of basic motives (drives) that are universal across individuals. Individual differences may be explained (in part) by the differences across this set of drives in drive strengths in different situations by different individuals. These drives, with their different strengths, lead to setting of different goals (as well as major cognitive parameters) by the metacognitive subsystem. Individual differences in terms of drive strengths are consequently reflected in the resulting goals (as well as major cognitive parameters). On the basis of the goals set (and the cognitive parameters chosen), an individual makes action decisions, within the action-centred subsystem (possibly in consultation with the non-action-centred subsystem). Thus their actions reflect their fundamental individual differences as well as situational factors as a result (Sun & Wilson, 2014). Their actions in turn affect the world in which they act.

Therefore, personality results from the complex interaction of many psychological entities, mechanisms, and processes, as well as their interaction with the world (Sun, 2016). Computational modelling and simulations with Clarion enabled us to see how exactly these entities, mechanisms and processes interact with each other. For computational details and simulations of personality, see Sun and Wilson (2014).

5.4. Emotion

Based on the Clarion framework, one hypothesis is that emotion is rooted in basic human motives (drives) and their possible fulfilment (Sun et al., 2016; Sun & Mathews, 2012). In this regard, some other researchers, for example, Smillie, Pickering, and Jackson (2006), Carver and Scheier (1990), and Ortony, Clore, and Collins (1988), also stressed the importance of motivation and expectation in emotion.

Thus emotions should be analysed in terms of their motivational underpinnings. For example, it may be hypothesised within the Clarion framework that the emotion of elation is related to positive reward (including unexpected positive reward) and also, to a lesser extent, 'expectation' of positive reward, on the basis of currently activated motivations (Ortony et al., 1988; Sun et al., 2016). The intensity of elation may be (in part) related to the strengths of approach-oriented drives in the motivational subsystem (Smillie et al., 2006; Sun et al., 2016).

For another example, the emotion of anxiety can be related to 'expectation' of negative feedback on the basis of currently activated motivations. The intensity of anxiety may be (in part) a function of the strengths of avoidance-oriented drives in the motivational subsystem (Sun et al., 2016). Smillie et al. (2006) specifically identified the link between the avoidance system and anxiety. Carver and Scheier (1990) also made related points. On the basis of the Clarion framework, similar descriptions can be applied to other basic emotions.

Within Clarion, emotional processes mainly occur in the bottom level (Sun et al., 2016; Sun & Mathews, 2012). That is, emotional processing is mostly implicit (Hudlicka & Fellous, 1996). It is closely tied to actions, that is, to the action-centred subsystem (e.g. Frijda, 1986). Processing within the action-centred subsystem occurs on the basis of drive activations from the motivational subsystem. Metacognitive control/regulation (by the metacognitive subsystem) occurs on the basis of drive activations as well. Explicit processes have an impact on emotion too, for example, through 'cognitive appraisal' (Frijda, 1986; Ortony et al., 1988). Thus, emotions involve dynamic interactions among various subsystems and processes within Clarion. See Wilson and Sun (2014) for simulations of emotion-related phenomena.

5.5. Morality

Within Clarion, there are two possible models of moral judgement. One corresponds to a reactive, situated view of the mind (e.g. Brooks, 1991). The other is more complex, and more reflective of the motivational views of the mind, as discussed earlier (Maslow, 1987; McDougall, 1908; Murray, 1938; Sun, 2009).

In the simpler model, decision-making happens in the action-centred subsystem, while the motivational subsystem only expresses a generic goal (e.g. to save life, but without considering the complexity of the matter, such as having to kill one in order to save many). Based on the generic goal, the action-centred subsystem undertakes the actual decision-making taking into consideration the full complexity of the matter (e.g. having to kill one in order to save many and whatever means necessary to do so). Its bottom-level implicit process makes decisions based on its reactive routines (i.e. 'moral instincts' in this case, biologically or socially formed), while its top-level explicit process performs explicit decision-making (e.g. explicit utilitarian calculation). This way of capturing the moral judgement may be justified based on the work of situated cognition theorists as discussed earlier (see, e.g. Brooks, 1991).

On the other hand, for the more complex motivationally based model, the metacognitive subsystem has to decide on a more specific goal (not just a vague desire) based on drive activations. In a way, it has to make a detailed 'decision' by considering the complexity of the matter (e.g. having to kill one in order to save many and the means of doing so). After a specific goal is generated, the action-centred subsystem takes that specific goal into consideration in directing the actual actions and generates the final action outputs. The action-centred subsystem, in this case, often does so in a more explicit and more deliberative way (e.g. by using the non-action-centred subsystem to perform detailed utilitarian calculation). The more complex model is justified based on the work concerning the complex dynamics of human motivation and its interaction with cognition; see, for example, Murray (1938), Maslow (1987), and McDougall (1908), as discussed earlier (see also Sun, 2009). For detailed simulations, see Sun (2013) and Bretz and Sun (2017).

6. General discussion

6.1. Contributions

Compared with other existent cognitive architectures, Clarion is unique in that it contains (1) sophisticated built-in motivational constructs, and (2) sophisticated built-in metacognitive constructs. Beyond these mechanisms, Clarion is also unique because of some generic characteristics such as (3) the separation of the two fundamental dichotomies: the dichotomy of implicit vs. explicit processes and the dichotomy of action-centred vs. non-action-centred processes, and (4) both bottom-up and top-down learning (from implicit learning to explicit learning, and vice versa; Sun, 2002, 2016). These (specific and generic) characteristics are not commonly found in other existing cognitive architectures (cf. Anderson & Lebiere, 1998; Laird, 2012). Nevertheless, we believe that these characteristics are crucial to cognitive architectures, as they capture important elements of the human mind.

Duality of implicit and explicit representation, and concomitant processes and mechanisms, are present in, and affect thereby, both primary control of action and also deeper layers of control, that is, motivational and metacognitive processes (within the MS and the MCS). The Clarion cognitive architecture captures this duality of representation, in both primary and deeper control, in contrast to other existing proposals of cognitive architectures.

In Clarion, built-in motivational and metacognitive representations, mechanisms and processes provide explanations for a variety of motivational and metacognitive phenomena. They provide deeper explanations of human behavioural propensities and variability than simple, direct control of action envisioned in many existing proposals of cognitive architectures. As all the previous arguments have shown, understanding such deeper layers of control, going beyond direct control to explore motivational and metacognitive processes, is important to understanding the human mind.

The inclusion of these processes in Clarion makes it possible to generate internal feedback (e.g. for reinforcement learning). Such feedback may be generated internally based on assessing the current state of the world (in relation to motivation and other information). Thus, an agent does not need an externally provided reinforcement signal as required by common reinforcement learning algorithms (Sutton and Barto, 1998). This way, the model captures human learning processes — in the natural world there is often no external feedback available but agents can learn nevertheless.

Furthermore, the built-in motivational and metacognitive mechanisms in Clarion encourage unified explanations (through simulations within a unified framework) of a variety of motivational and metacognitive data and phenomena, all within the Clarion cognitive architecture and based on the same set of built-in motivational and metacognitive primitives. Clearly, such unified explanations are an important goal of cognitive science.

The inclusion of these motivational and metacognitive processes in Clarion also makes it possible to account for complex structures of sequential behaviour (routines), such as persistence, interruptibility and so on, as discussed earlier (see Sun, 2009, 2016), which cannot be easily accounted for with a simple goal stack as used in some other proposed cognitive architectures.

6.2. Comparisons

There have been various notions proposed for analysing control of behaviour. For example, the idea of motor schema was elaborated by Arbib (1985), in which individual components of a schema that represented a skill unit were each represented by a lower-level schema. A related notion is 'script' (Schank & Abelson, 1977), which was proposed to represent stylised routine activities (such as going to a restaurant). Scripts were generally specified through sequences of actions, which might be altered under various conditions or combined with other activities. Contention scheduling (Norman & Shallice, 1986) has been implemented through a hierarchy of schemas. In comparison with Clarion, these proposals often ignore motivational origins of behaviour, and often lack detailed metacognitive monitoring and regulation (resulting in less flexibility in behaviour).

Existing proposals of cognitive architectures often do not have sufficiently detailed specifications of motivational aspects of human behaviour, nor much in the way of metacognition. Of course, some of the functionalities can conceivably be implemented in some existing cognitive architectures, but not all functionalities can be easily implemented this way. In addition, inclusion of specific subsystems for motivation and metacognition encourages more attention to be paid to these.

However, there have been a number of cognitive architectures proposed with a metacognitive component. They typically consist of three levels: the reactive level, the deliberate level and the metacognitive level (see, e.g. Gat, 1998; Sloman, 2000). Their three-level structuring is very similar to Clarion, which has two levels of implicit and explicit representations (in both the ACS and the NACS), plus the metacognitive representations (in the MCS). However, these models often do not have a motivational subsystem.

In contrast to other cognitive architectures reviewed above, Psi (Bach, 2009) deals with some motivational dynamics. In Psi, each goal-directed action has its source in a motive that connects a goal to an 'urge', which is related to a physiological, cognitive or social 'demand'. When a goal is reached, a demand may be fulfilled, which creates a pleasure signal that is used for learning. Although Psi is somewhat similar to Clarion (at an abstract level), it does not have as much empirical grounding: It has not been used extensively for modelling psychological processes in accordance with empirical data.

A quick comparison with specialised models of motivation is in order. McDougall (1908) proposed a number of specific 'instincts' to explain behaviour. Freud (1915) introduced the concept of Trieb. Tolman (1932) and Hull (1951) utilised the notion of drive. Lewin (1936) introduced the concept of need and two categories of needs (biological and social). These models (and others reviewed earlier) introduced useful notions into the discussion of motivation. However, most of these models did not provide detailed process-based accounts of motivation.

Sloman (1986) hypothesised a taxonomy of motives and their working in terms of goal setting. His models generally focused on explicit representations, not complex implicit motivational dynamics. As our preceding arguments showed, implicit drives and their complex interactions are important factors in human motivation. For other computational perspectives on motivation, see Merrick and Maher (2009) and Baldassarre and Mirolli (2013).

Let us turn to specialised models of metacognition. Nelson and Narens (1990) divided cognitive processes into a meta-level and an object level, whereby the meta-level had a model of the object level informed by monitoring activities, and exerted control on the object-level processes. Reder and Schunn's (1996) model was based on spreading activation among related items in a semantic memory, and involved only implicit processes. Reder's (1987) hypothesis that there were two separate processes (one rapid and the other slow) was, however, similar to our distinction of the two types of processes (implicit and explicit). Norman and Shallice's (1986) model, as mentioned before, was akin to our model, but we address more complex forms of implicit and explicit metacognitive mechanisms and processes and allow intermeshed metacognitive and regular processes. For more computational perspectives on metacognition, see Caro, Josyula, Cox, and Jiménez (2014). In all, Clarion provides a different perspective based on a two-level, dual-representational framework.

6.3. Concluding remarks

In this article, arguments were presented regarding deeper (motivational and metacognitive) control of human behaviour. These arguments led to a cognitive architecture Clarion. Clarion takes into account motivational, cognitive, and metacognitive processes and their interactions. With these, Clarion has something to contribute to cognitive modelling – it aims to capture the interaction among motivational, metacognitive and cognitive processes, and to explain their functioning in concrete computational terms. This work constitutes an attempt at a detailed theory of these processes and their interactions. As such, it is subject to revision, refinement and further development.

The upshot is that it is necessary to incorporate all of the essential processes, including motivational and metacognitive processes, into a comprehensive theory of the mind in the form of a computational cognitive architecture. Short of that, our understanding of cognition can only be partial and incomplete, and the progress of cognitive science may be hampered.

Note

1. In a way, this is akin to the BDI framework, but in a more nuanced, more psychologically realistic way (Sun, 2016).

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported in part by the Army Research Institute for the Behavioural and Social Sciences [grant number W911NF-17-1-0236].

References

Aarts, H., & Hassin, R. (2005). Automatic goal inference and contagion. In J. Forgas, K. Williams, & S. Laham (Eds.), *Social Motivation: Conscious and Unconscious Processes*. New York, NY: Cambridge University Press.

Adams, H., Wright, L., & Lohr, B. (1996). Is homophobia associated with homosexual arousal? *Journal of Abnormal Psychology, 105*(3), 440–445.

Anderson, J., & Lebiere, C. (1998). *The Atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Arbib, M. (1985). Schemas for the temporal organization of behavior. *Human Neurobiology, 4*, 63–72.

Atkinson, J. (1964). *An introduction to motivation*. Princeton, NJ: Van Nostrand.

Bach, J. (2009). *Principles of synthetic intelligence PSI: An architecture of motivated cognition*. New York, NY: Oxford University Press.

Baldassarre, G., & Mirolli, M. (2013). *Intrinsically motivated learning in natural and artificial systems*. Berlin: Springer-Verlag.

Bargh, J. (1997). Advances in social cognition. In R. Wyer (Ed.), *The Automaticity of Everyday Life* (pp. 1–61). Mahwah, NJ: Erlbaum.

Bensen, D. (1994). *The neurology of knowledge*. New York, NY: Oxford University.

Berridge, K., & Schulkin, J. (1989). Palatability shift of a salt-associated incentive during sodium depletion. *Quarterly Journal of Experimental Psychology: Comparative and Physiological Psychology, 41*, 121–138.

Bickhard, M. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence,* 285–333.

Bretz, S., & Sun, R. (in press). Two models of moral judgment. *Cognitive Science*.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence, 47*, 139–160.

Bruner, J., Goodnow, J., & Austin, J. (1956). *A study of thinking*. New York, NY: Wiley.

Burns, T., & Roszkowska, E. (2006). Social judgment in multi-agent systems. In R. Sun (Ed.), *Cognition and Multi-Agent Interaction: Integrating Cognitive Modeling and Social Simulation*. New York, NY: Cambridge University Press.

Busemeyer, J., Townsend, J. T., & Stout, J. C. (2002). Motivational underpinnings of utility in decision making: Decision field theory analysis of deprivation and satiation. In S. Moore (Ed.), *Emotional Cognition*. Amsterdam: John Benjamins.

Carbonell, J. G., Knoblock, C. A., & Minton, S. (1991). PRODIGY: An integrated architecture for planning and learning. In K. VanLehn (Ed.), *Architectures for Intelligence* (pp. 241–278). Hillsdale, NJ: Lawrence Erlbaum Associates.

Caro, M. F., Josyula, D. P., Cox, M. T., & Jiménez, J. A. (2014). Design and validation of a metamodel for metacognition support in artificial intelligent systems. *Biologically Inspired Cognitive Architectures, 9*, 82–104.

Carver, C., & Scheier, M. (1990). Origin and functions of positive and negative affect: A control – process view. *Psychological Review, 97*, 19–35.

Castelfranchi, C. (2001). The theory of social functions: Challenges for computational social science and multi-agent learning. *Cognitive systems research, special issue on the multi-disciplinary studies of multi-agent learning*. R. Sun (Ed.), *2*(1), 5–38.

Chaiken, S. & Trope, Y. (Eds.). (1999). *dual process theories in social psychology*. New York, NY: Guilford Press.

Chen, S., Shechter, D., & Chaiken, S. (1996). Getting at the truth or getting along: Accuracy – vs. impression-motivated heuristic and systematic processing. *Journal of Personality and Social Psychology, 71*(2), 262–275.

Clancey, W. (1997). *Situated cognition: On human knowledge and computer representations*. Cambridge: Cambridge University Press.

Clark, A., & Karmiloff-Smith, A. (1993). The Cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language, 8*(4), 487–519.

Cleeremans, A. (1997). Principles for implicit learning. In D. Berry (Ed.), *How implicit is implicit learning?* (pp. 195–234). Oxford: Oxford University Press.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences, 2*(10), 406–416.

Dai, D. Y., & Sternberg, R. J. (2004). *Motivation, emotion, and cognition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Darling, S., Sala, S., Gray, C., Trivelli, C., Mazzoni, G., & Nelson, T. (1998). Putative functions of the prefrontal cortex. In *Metacognition and Cognitive Neuropsychology*. Mahwah, NJ: Erlbaum.

Dayan, P. (2001). *Motivated reinforcement learning. Neural information processing*. Cambridge, MA: MIT Press.

Deci, E. & Staub, E. (1980). Intrinsic motivation and personality. In *Personality: Basic issues and current research* (pp. 35–80). Englewood Cliffs, NJ: Prentice Hall.

Derryberry, D., & Tucker, D. (1994). Motivating the focus of attention. In P. Miedenthal & S. Kitayama (Eds.), *Heart's Eye: Emotional Influences on Perception and Attention* (pp. 167–196). San Diego, CA: Academic Press.

Dethier, V. (1966). Insects and the concept of motivation. *Nebraska Symposium on Motivation, 14*, 105–136.

Eccles, J., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132.

Epstein, A. (1982). Instinct and motivation as explanations for complex behavior. In D. W. Pfaff (Ed.), *The Physiological Mechanisms of Motivation*. Berlin: Springer-Verlag.

Evans, J. & Frankish, K. (Eds.). (2009). *In two minds: Dual processes and beyond*. Oxford: Oxford University Press.

Flavell, J. (1976). Metacognitive aspects of problem solving. In B. Resnick (Ed.), *The Nature of Intelligence*. Hillsdale, NJ: Erlbaum.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. In S. Pinker & J. Mehler (Eds.), *Connections and Symbols*. Cambridge, MA: MIT Press.

Forgas, J., Williams, K., & Laham, S. (Eds.). (2005). *Social motivation: Conscious and unconscious processes*. New York, NY: Cambridge University Press.

Forrest-Pressley, D. L. & Waller, T. G. (1984). *Cognition, metacognition, and reading*. New York, NY: Springer-Verlag.

Freud, S. (1915). *A general introduction to psychoanalysis*. New York, NY: Washington Square.

Frijda, N. (1986). *The emotion*. New York, NY: Cambridge University Press.

Gat, E. (1998). On three-layered architecture. In D. Kortenkamp, R. Bonasso, & R. Murphy (Eds.), *Artificial Intelligence and Mobile Robots*. Menlo Park, CA: AAAI Press.

Gentner, D., & Collins, A. (1981). Studies of inference from lack of knowledge. *Memory and Cognition, 9*, 434–443.

Grafman, J. (1994). Alternative frameworks for the conceptualization of frontal lobe functions. In F. Boller & J. Grafman (Eds.), *Handbook of Neuropsychology*, Vol. 9 (pp. 187–202). Amsterdam: Elsevier.

Grossberg, S. (1971). On the dynamics of operant conditioning. *Journal of Theoretical Biology, 33*, 225–255.

Hayes-Roth, F., Waterman, D. A., & Lenat, D. B. (Eds.). (1983). *Building expert systems*. Reading, MA: Addison- Wesley.

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

Herrstein, R. (1977). The evolution of behaviorism. *American Psychologist, 32*, 593–603.

Hertel, G., Kerr, N., & Messe, L. (2000). Motivation gains in performance groups: Paradigmatic and theoretical development on the Kohler effect. *Journal of Personality and Social Psychology, 79*(4), 580–601.

Hing, L., Chung-Yan, G., Grunfeld, R., Robichaud, L., & Zanna, M. (2005). Exploring the discrepancy between implicit and explicit prejudice. In J. Forgas, K. Williams, & S. Laham (Eds.), *Social Motivation: Conscious and Unconscious Processes*. New York, NY: Cambridge University Press.

Hintzman, D. (1990). Human learning and memory: Connections and dissociations. *Annual Review of Psychology, 41*, 109–139.

Hudlicka, E., & Fellous, J. (1996). *Reviews of computational models of emotion*. Manuscript.

Hull, C. (1951). *Essentials of behavior*. New Haven, CT: Yale University Press.

Iyengar, S., & Lepper, M. (1999). Rethinking the value of choice: A cultural perspective on intrinsic motivation. *Journal of Personality and Social Psychology, 76*(3), 349–366.

Kanfer, R., & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology, 74*(4), 657–690.

Karmiloff-Smith, A. (1986). From meta-processes to conscious access: Evidence from children's metalinguistic and repair data. *Cognition, 23*, 95–147.

Kernis, M., & Goldman, B. (2005). Authenticity, social motivation, and psychological adjustment. In J. Forgas, K. Williams, & S. Laham (Eds.), *Social Motivation: Conscious and Unconscious Processes*. New York, NY: Cambridge University Press.

Koriat, A., & Goldsmith, M. (1998). The role of metacognitive processes in the regulation of memory performance. In G. Mazzoni & T. Nelson (Eds.), *Metacognition and Cognitive Neuropsychology*. Mahwah, NJ: Erlbaum.

Laird, J. (2012). *The soar cognitive architecture*. Cambridge, MA: MIT Press.

Lashley, S. (1951). The problem of serial order in behavior. In L. Jeffress (Ed.), *Cerebral Mechanisms in Behavior*. New York, NY: Wiley.

Latham, G., & Pinder, C. (2005). Work motivation theory and research at the dawn of the twenty-first century. *Annual Review of Psychology, 56*, 485–516.

Leven, S., & Levine, D. (1996). Multiattribute decision making in context: A dynamic neural network methodology. *Cognitive Science, 20*, 271–299.

Lewin, K. (1936). *Principles of topological psychology*. New York, NY: McGraw-Hill.

Logan, G. (2003). Executive control of thought and action: In search of the wild homunculus. *Current Directions in Psychological Science,* 45–48.

Logan, G., & Gordon, R. (2001). Executive control of visual attention in dual-task situations. *Psychological Review, 108*, 393–434.

Lorenz, K. (1950). The comparative method in studying innate behavior patterns. *Symposia of the Society for Experimental Biology* (Physiological Mechanisms in Animal Behavior), *4*, 221–254.

Luria, A. R. (1966). *Higher cortical functions in man*. New York, NY: Basic Books.

Maheswaran, D., & Chaiken, S. (1991). Promoting systematic processing in low-motivation settings: Effect of incongruent information on processing and judgment. *Journal of Personality and Social Psychology, 61*(1), 13–25.

Maner, J., Kenrick, D., Becker, D., Robertson, T., Hofer, B., Neuberg, S., … Schaller, M. (2005). Functional projection: How fundamental social motives can bias interpersonal perception. *Journal of Personality and Social Psychology, 88*(1), 63–78.

Maslow, A. (1987). *Motivation and personality* (3rd ed.). New York, NY: Harper and Row.

Mazzoni, G. & Nelson, T. (Eds.). (1998). *Metacognition and cognitive neuropsychology*. Mahwah, NJ: Erlbaum.

McClelland, D. (1951). *Personality*. New York, NY: Dryden Press.

McDougall, W. (1908). *Introduction to social psychology*. London: Methuen.

McFarland, D. (1989). *Problems of animal behaviour*. Singapore: Longman Publishing.

Merrick, E., & Maher, M. L. (2009). *Motivated reinforcement learning*. Berlin: Springer-Verlag.

Metcalfe, J. (1986). Dynamic metacognitive monitoring during problem solving. *Journal of Experimental Psychology: Learning, Memory and Cognition, 12*, 623–634.

Metcalfe, J. & Shimamura, A. (Eds.). (1994). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press.

Murray, H. A. (1938). *Explorations in personality*. New York, NY: Oxford University Press.

Narens, T., Graf, A., & Nelson, T. (1996). Metacognitive aspects of implicit/explicit memory. In L. Reder (Ed.), *Implicit Memory and Metacognition*. Mahwah, NJ: Erlbaum.

Nelson, T., & Narens, L. (1990). Meta-memory: A theoretical treatment and new findings. In G. Bower (Ed.), *The Psychology of Learning and Motivation*, Vol. 26 (pp. 125–140). New York, NY: Academic Press.

Neuberg, S., Kenrick, D., Manor, J., & Shaller, M. (2005). From evolved motives to everyday mentation. In J. Forgas, K. Williams, & S. Laham (Eds.), *Social motivation: Conscious and unconscious processes*. New York, NY: Cambridge University Press.

Norman, D., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In G. Schwartz & D. Shapiro (Eds.), *Consciousness and self regulation: Advances in research and theory*, Vol. 4 (pp. 1–18). New York, NY: Plenum.

Norton, M., Vandello, J., & Darley, J. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology, 87*(6), 817–831.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions*. New York, NY: Cambridge University Press.

Pennisi, E. (2005). How did cooperative behavior evolve? *Science, 309*, 93.

Pintrich, P., Marx, R., & Boyle, R. (2003). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research, 63*(2), 167–199.

Read, S. J., Monroe, B. M., Brownstein, A. L., Yang, Y., Chopra, G., & Miller, L. C. (2010). A neural network model of the structure and dynamics of human personality. *Psychological Review, 117*(1), 61–92.

Reber, A. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118*(3), 219–235.

Reder, L. (1987). Strategy selection in question answering. *Cognitive Psychology, 19*, 111–138.

Reder, L. (Ed.). (1996). *Implicit memory and metacognition*. Mahwah, NJ: Erlbaum.

Reder, L., & Schunn, C. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. Reder (Ed.), *Implicit Memory and Metacognition*. NJ, Mahwah: Erlbaum.

Reiss, S. (2004). Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. *Review of General Psychology, 8*(3), 179–193.

Ricoeur, P. (1981). The metaphorical process as cognition, imagination, and feeling. In M. Johnson (Ed.), *Philosophical Perspectives on Metaphor*. MN: University of Minnesota Press, Minneapolis.

Rumelhart, D. J., McClelland, D. C., & The PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructures of cognition*. Cambridge, MA: MIT Press.

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

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals, and understanding*. Mahwah, NJ: Erlbaum Associates.

Scheutz, M., & Sloman, A. (2001). *Affect and agent control: Experiments with simple affective states. IAT'01*. Singapore: World Scientific.

Schneider, W. (1998). The development of procedural metamemory in childhood and adolescence. In G. Mazzoni & T. Nelson (Eds.), *Metacognition and Cognitive Neuropsychology*. Mahwah, NJ: Erlbaum.

Schooler, J., & Schreiber, C. (2005). To know or not to know. In J. Forgas, K. Williams, &. S. Laham (Eds.), *Social Motivation: Conscious and Unconscious Processes*. New York, NY: Cambridge University Press.

Seger, C. (1994). Implicit learning. *Psychological Bulletin., 115*(2), 163–196.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Simon, H. (1967). Motivational and emotional controls of cognition. *Psychological Review, 74*, 29–39.

Sloman, A. (1986). *Motives mechanisms and emotions*. Cognitive science research papers CSRP 062. Falmer: University of Sussex.

Sloman, A. (2000). Architectural requirements for human-like agents both natural and artificial. In K. Dautenhahn (Ed.), *Human Cognition and Social Agent Technology*. Amsterdam: John Benjamins.

Sloman, A., & Chrisley, R. (2005). More things than are dreamt of in your biology: Information processing in biologically-inspired robots. *Cognitive Systems Research, 6*(2), 145–174.

Smillie, L. D., Pickering, A. D., & Jackson, C. J. (2006). The new reinforcement sensitivity theory: Implications for personality measurement. *Personality and Social Psychology Review, 10*, 320–335.

Snyder, M., & Stukas, A. (1999). Interpersonal processes: The interplay of cognitive, motivational, and behavioral activities in social interaction. *Annual Review of Psychology, 50*, 273–303.

Stout, J., Busemeyer, J. R., Lin, A., Grant, S. R., & Bonson, K. R. (2004). Cognitive modeling analysis of the decision-making processes used by cocaine abusers. *Psychonomic Bulletin and Review., 11*(4), 742–747.

Strack, F., & Deutsch, R. (2005). Reflection and impulse as determinants of conscious and unconscious motivation. In J. Forgas, K. Williams, & S. Laham (Eds.), *Social Motivation: Conscious and Unconscious Processes*. New York, NY: Cambridge University Press.

Stuss, D., & Benson, D. (1986). *The frontal lobe*. New York, NY: Raven.

Sun, R. (1994). *Integrating rules and connectionism for robust commonsense reasoning*. New York, NY: Wiley.

Sun, R. (2001). Meta-learning in multi-agent systems. In N. Zhong, J. Liu, S. Ohsuga, & J. Bradshaw (Eds.), *Intelligent Agent Technology: Systems, Methodologies, and Tools*. Singapore: World Scientific.

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

Sun, R. (2009). Motivational representations within a computational cognitive architecture. *Cognitive Computation, 1*(1), 91–103.

Sun, R. (2013). Moral judgment, human motivation, and neural networks. *Cognitive Computation, 5*(4), 566–579.

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

Sun, R., & Mathews, R. (2012). Implicit cognition, emotion, and meta-cognitive control. *Mind and Society, 11*(1), 107–119.

Sun, R., & Sessions, C. (2000). Self-segmentation of sequences: Automatic formation of hierarchies of sequential behaviors. *IEEE Transactions on Systems, Man, and Cybernetics: Part B, Cybernetics, 30*(3), 403–418.

Sun, R., & Wilson, N. (2014). A model of personality should be a cognitive architecture itself. *Cognitive Systems Research, 29–30*, 1–30.

Sun, R., Merrill, E., & Peterson, T. (2001). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science, 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. *Psychological Review, 112*(1), 159–192.

Sun, R., Zhang, X., & Mathews, R. (2006). Modeling meta-cognition in a cognitive architecture. *Cognitive Systems Research, 7*(4), 327–338.

Sun, R., Wilson, N., & Lynch, M. (2016). Emotion: A unified mechanistic interpretation from a cognitive architecture. *Cognitive Computation, 8*(1), 1–14.

Sutton, R. & Barto, A. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Tauer, J., & Harackiewicz, J. (2004). The effects of cooperation and competition on intrinsic motivation and performance. *Journal of Personality and Social Psychology, 86*(6), 849–861.

Tinbergen, N. (1951). *The study of instinct*. London: Oxford University Press.

Toates, F. (1986). *Motivational systems*. Cambridge: Cambridge University Press.

Tolman, E. C. (1932). *Purposive behavior in animals and men*. New York, NY: Appleton-Century-Crofts.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

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

Umilta, C., & Stablum, F. (1998). Control processes explored by the study of closed-head-injury patients. In G. Mazzoni & T. Nelson (Eds.), *Metacognition and Cognitive Neuropsychology*. Mahwah, NJ: Erlbaum.

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

Wegner, D. (1994). Ironic processes of mental control. *Psychological Review, 101*, 34–52.

Weiner, B. (1992). *Human motivation: Metaphors, theories, and research*. Newbury Park, CA: Sage.

Wilson, N., & Sun, R. (2014). Coping with bullying: A computational emotion-theoretic account. In *Proceedings of the Annual Conference of Cognitive Science Society* (pp. 3119–3124). Quebec, Canada. Austin, TX. Published by Cognitive Science Society.

Wilson, N., Sun, R., & Mathews, R. (2009). Performance degradation under pressure. *Neural Networks, 22*, 502–508.

Woike, B. (1995). Most memorable experiences: Evidence for a link between implicit and explicit motives and social cognitive processes in everyday life. *Journal of Personality and Social Psychology, 68*, 1081–1091.

Wood, W., & Quinn, J. (2005). Habits and the structure of motivation in everyday life. In J. Forgas, K. Williams, & S. Laham (Eds.), *Social motivation: Conscious and unconscious processes*. New York, NY: Cambridge University Press.