

DOI: 10.20098/j.cnki.1001-5132.2024.1014

Computational cognitive architectures, dual-process theories, and LLMs

SUN Ron

(Cognitive Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180, USA)

Abstract: This paper introduces a computational cognitive architecture that serves as a comprehensive computational theory of the human mind, from cognitive science and computational psychology. The cognitive architecture (named Clarion) has been justified by, and validated against, psychological data, findings, and theoretical constructs. One important theoretical background for it is the dual-process theories, which led to its overall two-level structuring in a hybrid neuro-symbolic way. Furthermore, given the recent advances in AI and computing technology, LLMs are being incorporated into the model to better capture human intuition and instinct (and implicit processes in general), in order to further enhance Clarion. Integrating Clarion and LLMs can also help to develop AI systems that are more capable, more reliable, and more human-like. Overall, the paper advocates a multidisciplinary approach towards developing better models for cognitive science and for AI.

Key words: cognitive architecture; LLM; dual process; psychology; AI

CLC number: TP18

Document code: A

Article ID: 1001-5132 (2025) 02-0017-12

1 Introduction

A computational cognitive architecture, generally speaking, is a comprehensive computational theory of the human mind, from cognitive science and computational psychology. Given its computational nature, it can be detailed yet precise, shedding light on the human mental processes beyond what other approaches (e.g., verbal-conceptual theories, or pure mathematical theories) can^[1]. One particular example of a computational cognitive architecture is Clarion, which has been developed since the 1990s (see, e.g., [2-3]). It has been extensively justified by, and validated against, psychological data, findings, and theoretical constructs, and accounts for a wide range of psychological data and

findings, more so than other existing cognitive architectures.

One important theoretical background for Clarion is the *dual-process theories* from the psychology and philosophy literatures, which posit the co-existence of implicit processes (including, e.g., intuition and instinct) and explicit processes. These ideas led to the overall two-level structuring of Clarion in a hybrid neuro-symbolic way^[4-5].

Furthermore, given the recent advances in AI and computing technology, LLMs are being incorporated into Clarion to better capture human intuition, instinct, and other implicit processes, in order to further enhance Clarion. Integrating Clarion and LLMs can also help to develop AI systems that are more capable, more

powerful, more reliable, and more human-like^[6]. Overall, the present article advocates a multidisciplinary approach towards developing better models for both cognitive science and AI.

In the following, this long-standing research program will be sketched. Note that, due to space limitation, the amount of detail has to be limited in the present article. First, in Section 2, dual-process theories and neuro-symbolic models are briefly explained. Then, in Section 3, the computational cognitive architecture Clarion is introduced. In Section 4, a more detailed discussion ensues that addresses how LLMs can be integrated into Clarion. Some concluding remarks in Section 5 end the paper.

2 Dual-process theories and neuro-symbolic models

Dual-process theories have been steadily gaining attention in recent decades, in psychology, philosophy, and many other fields. The distinction between implicit processes (also termed System 1 or simply “intuition”) and explicit processes (also termed System 2 or “reason”) has been argued by many since the 1980s (e.g., [4,7-8]) and popularized later by Kahneman^[9].

Generally speaking, explicit processes are relatively easily accessible to consciousness, while implicit processes are less so. Explicit processing may be described as symbolic and rule-based to some significant extent, while implicit processing is more “associative” and “holistic”^[4]. Empirical evidence and theoretical analysis in support of these points can be found in the literature (e.g., in the work cited above). Work on computational cognitive architectures has also demonstrated computationally the importance of dual processes (e.g., [10-11]; more on this later).

It has been argued a long time ago (e.g., [4]) as well as very recently (e.g., [5]) that hybrid neuro-symbolic systems may be necessary to address the dual processes of the human mind. For example, the neural (subsymbolic) side of a hybrid system may address low-

level sensory-motor processes and other unconscious (implicit) processes. The symbolic side can perform high-level conscious reasoning, explicit planning, and deliberative reflection. These two sides may be distinct structurally, representationally, and mechanistically, but work together in a synergistic way. Such neuro-symbolic models for tackling dual mental processes have been proposed since the 1990s, following the initial rise of dual-process theories a little earlier. Theoretical arguments for them have also emerged over the years. Furthermore, it has been emphasized that such systems should be rigorously structured in a cognitively (psychologically) motivated and justified way, based on the structures, mechanisms, representations, and contents of human mental processes^[5].

Overall, it has become increasingly evident that dual processes are a key feature of the human mind^[4,7-9], and hybrid neuro-symbolic systems are a natural fit for them.

3 Computational cognitive architectures

3.1 Why cognitive architectures?

First, computational psychology, as mentioned earlier, is concerned with the understanding of the human mind through developing computational models in a psychologically rigorous way^[1]. These models embody descriptions of psychological functioning in computer algorithms and programs. That is, they impute computational processes onto psychological functions. Thereby they lead to runnable computer programs. Detailed simulations and other operations can be conducted based on these models. Given the complexity of the human mind and the resulting behavioral complexity, precise, mechanistic, process-oriented computational models are necessary to explicate the intricate details of the human mind. In this regard, a computational model may constitute a theory by itself. This approach offers flexibility and expressive power that no other approach can match. It provides a variety

of modeling paradigms, methodologies, and techniques. It also supports practical applications of cognitive/psychological theories^[12].

Within this field, a *computational cognitive architecture* is an empirically (psychologically) grounded, comprehensive computational framework for understanding and potentially capturing computationally the human mind (e.g., [3]). Like architecture of buildings (which, as we know, includes frames, foundations, floors, roofs, and other essential aspects, but not chairs, tables, TVs, etc.), a cognitive architecture includes those and only those aspects within the mind that are structural, essential, or stable.

The advantage of such a comprehensive model is that it provides an essential framework that facilitates more detailed exploration of the mind. The many assumptions that it embodies may be based on empirical data, philosophical analysis, or computationally inspired hypotheses. Through embodying these assumptions, a cognitive architecture narrows down possibilities and provides scaffolding structures. Note that here only cognitive architectures that are meant to be rigorous models of the human mind (i.e., have been validated empirically) are relevant.

In such models, dual processes can be considered and mechanistically accounted for. Dual-process computational cognitive architectures explore psychological nuances and mechanistic details of implicit versus explicit psychological processes.

3.2 Clarion

The Clarion cognitive architecture will be used as an example here^[2-3]. It is a two-level, dual-process, dual-representational, and neuro-symbolic cognitive architecture. Clarion covers a wider range of psychological phenomena than any other cognitive architecture. Moreover, psychological processes in Clarion are motivationally grounded.

Clarion consists of four major subsystems: (1) the action-centered subsystem (ACS) for dealing with actions (internal or external), involving procedural (i.e., action-oriented) knowledge, (2) the non-action-centered

subsystem (NACS) for reasoning and memory involving declarative (i.e., factual) knowledge, as directed by the ACS/MCS, (3) the motivational subsystem (MS) for providing motivation (drives and goals), and (4) the metacognitive subsystem (MCS) for regulating other subsystems (on the basis of the MS and so on).

Each of these subsystems consists of two “levels”. The top level carries out explicit processes, while the bottom level carries out implicit processes^[2,4,8]. Computationally, one is symbolic and the other neural/subsymbolic (e.g., a backpropagation neural network). The two types of representations are inter-connected: The symbolic representations at the top level, which are in the forms of “chunk” nodes (each representing a concept, defined by a set of microfeatures at the bottom level) and rules connecting the chunk nodes, are linked to corresponding neural representations at the bottom level. The two processes thus interact to generate combined outcomes. See Fig. 1.

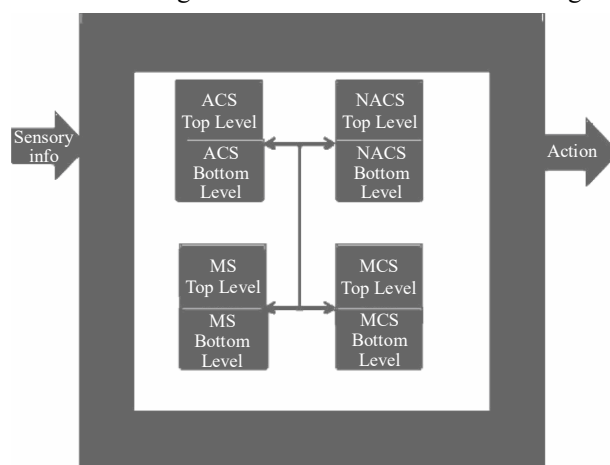


Fig. 1 The original Clarion cognitive architecture with its four major subsystems

In Clarion, detailed processes of implicit-explicit interaction have been elucidated mechanistically (i.e., computationally, in the ACS and the NACS, respectively^[3]) and validated psychologically (through modeling and simulation of empirical psychological data; see, e.g., [11]). The interaction is ever-present in Clarion, leading to complex dynamics, which contrasts with alternative theories such as “default-intervention” (cf. [7]).

In Clarion, one essential form of implicit-explicit interaction is what has been termed *top-down activation* (initiation of implicit processes at the bottom level by explicit processes at the top level), while the other is *bottom-up activation* (activation of symbolic representations at the top level by implicit processes at the bottom level). In this way, the two levels affect each other. For instance, implicit processes may be guided by explicit processes (e.g., by explicit instructions); alternatively, results of implicit processes may be verified, refined, rectified, or explained by explicit processes. Note that the simplest way of integrating the outcomes across levels is combining the outputs of the two levels through a weighted sum, whereby the weights are determined based on relative performance of each level.

In relation to learning, a form of implicit-explicit interaction in Clarion is *top-down learning* (taking explicit knowledge and assimilating it into implicit processes), and another is *bottom-up learning* (learning implicit knowledge first and then learning explicit knowledge on that basis). See Sun^[3] for technical details of these forms of learning, which (along with other forms of learning) enable incremental, continuous, real-time learning in Clarion from situations encountered (more on this later).

In Clarion, across different subsystems, there are a variety of memories: procedural memory, semantic memory, episodic memory, working memory, and so on, often in both implicit and explicit forms. These types of memory are crucially important for human cognition, as has been demonstrated time and again in empirical psychological research (e.g., [13-15]). They are also important to Clarion in capturing human cognition^[3]. Among them, procedural memory is located within the ACS (concerning action selection), and semantic and episodic memories are in the NACS (for reasoning, etc.). In addition, working memory is standalone, serving as supplemental input.

A possible flow of information across subsystems is roughly as follows: First, within the MS, situational

inputs trigger internal motives, termed *drives*, based on both internal propensities and situational inputs. Then, a goal (an intention for action) within the MS is selected (by the MCS) on the basis of activated motives/drives (for the sake of satisfying them). Action/meta-action selection occurs, within the ACS/MCS, on the basis of the goal selected and the situational inputs (to maximize the satisfaction of the motives/drives). Note that some of the selected actions/meta-actions are for internal operations, including directing reasoning by the NACS; thus, the NACS is under the control of the ACS/MCS.

For any organism, its main motive is sustaining the individual life. Beyond that, another motive is sustaining the species (through reproduction). However, a number of other essential motives (including socially oriented ones) also exist (depending on species, of course). Humans have more complex and more sophisticated motivational structures (comprising drives, goals, and so on). Cognition has evolved to serve them. Human motives have been explored by social psychologists (as well as ethologists). From Murray^[16] to Reiss^[17], essential human motives, which are considered fundamentally important, have been identified through empirical and theoretical work. Correspondingly, they are also a fundamentally important part of Clarion, including: *achievement, affiliation, power, autonomy, deference, nurturance*, and other drives (within the MS; for the complete list, see [3]). Goals (which direct actions) are determined on the basis of drives, together constituting dual processes of motivation.

Clarion has thus far accounted for empirical psychological findings in a wide range of domains (see [10-11,18]; etc.). For example, a variety of well-known skill learning tasks have been simulated and explained using Clarion, some of which are highly implicit, while others involve more explicit processes. Simulations have also been done with reasoning tasks, metacognitive and motivational tasks, and social simulation tasks. Thus, Clarion has been well validated^[3]. It constitutes a comprehensive theory of the mind. Due to space limitation, for further details, see [2-3].

4 Integrating LLMs into Clarion

4.1 Strengths and shortcomings of current LLMs

On the other hand, LLMs have achieved spectacular successes in recent years (e.g., in natural language generation or human-level test performance). These systems have found their way into large-scale, real-world applications, dealing with texts, images, voices, and so on.

However, at the same time, compared with humans, current LLMs suffer from some shortcomings, including limited abstract reasoning or planning capabilities, limited memory, lack of autonomy, limited reliability, and so on. For instance, such systems often produce content that is meaningless or false. They may be inconsistent with human preferences, deviating from human values or ethics. When faced with complex problems, LLMs may have difficulties achieving human-level performance using various cognitive strategies that humans possess.

These shortcomings often stem from not taking human psychology into serious enough consideration. Thus the understanding of the human mind may help in this regard. Future systems achieving human-level performance should, almost by necessity, draw inspirations from human psychology and models of the human mind (e.g., from computational psychology). In the ongoing discourse regarding the limits of current LLMs, the debate is often between those who think that more data alone can provide the solution and those who favor incorporating additional methods. From the above perspective, the latter position seems more plausible, but it may need to be guided by empirically grounded understanding of the human mind^[6].

One aspect of the human mind is, of course, dual processes, which was discussed earlier and well developed in Clarion. Another aspect is essential human motives, which are key elements of the human mind^[16-17] and the basis of behavior regulation^[19]. This aspect has been well developed in Clarion too. Of course, there are many other aspects of the human mind embodied by

Clarion, as will be explicated later. Therefore, Clarion seems to be a suitable framework for enhancing LLMs^[6,20].

From a different perspective, some computational cognitive architectures have been in development since at least the 1980s. While they are theoretical models of the human mind, their actual capabilities have not been keeping up with the advances in technology. Therefore, computational cognitive architectures need to incorporate new advances in AI and in computing technology, more readily and more seamlessly. LLMs may serve as the key to this endeavor due to their capabilities and potentials^[6].

4.2 What can LLMs do?

In characterizing LLMs, one may consider human intuition and instinct^[18,21-22], besides implicit processes in general, of course. Here, intuition refers to understanding without the involvement of explicit reason (regardless of whether it is immediate or slow in emergence; [23]), while instinct refers to relatively fixed patterns of responses to environmental stimuli without involving explicit reason^[18,24]. In this regard, one may argue that LLMs correspond roughly to human intuition and instinct (setting aside perception, motor control, language, and other peripheral implicit processes, as well as motivation, for the time being; [6]).

Why should LLMs correspond to intuition and instinct? What characteristics of LLMs make them correspond to intuition and instinct? Let us look into intuition first (see [6] for more details).

Human intuition means unconscious (implicit) processes^[7-9]. However, one may view intuition as a form of reasoning: Reasoning encompasses explicit processes on the one hand, and implicit processes (intuition) on the other^[4], both guided by one's motives and goals. In fact, intuition, as well as insight resulting from intuition, is arguably indispensable elements in human thinking — They may supplement and guide explicit reasoning^[10]. Prior empirical work demonstrating the above can be found in [7-8] and so on, while theoretical work in this regard includes [10,23] and others.

Intuition might often be, though not necessarily always, beyond what language can express; at least, the process leading up to intuition and insight is often beyond language^[4,22]. However, intuition is at least partially expressible through language. Given an extremely large amount of linguistic data, it is likely that human intuition about a vast array of subjects may be contained in (or embodied by) it. With such data, it may be argued that LLMs resulting from training can gain a large range of intuition, somewhat comparable to human intuition. That is, although an individual piece of intuition might not be captured by an individual sentence, a vast number of sentences from a vast number of different sources can conceivably capture a fair amount of intuition collectively. Furthermore, human intuition results from and relies on diverse (likely implicit) commonsense knowledge about how the world works, including, for example, naive physics, folk psychology, folk sociology, and so on. This is what an extremely large amount of linguistic training data from diverse sources can provide. On top of that, the training process with gradual numerical weight adjustments (in Transformers, as the basis of LLMs) leads to generalization, which in turn leads to better intuition of an even broader scope. Thus, conceivably, LLMs can embody human intuition as gleaned and generalized from a large volume of data^[6].

In humans, intuition is often obtained through experience, especially a large amount of often repeated experience. A large amount of repeated experience is exactly what the training of LLMs with a large amount of data provides. Through the experience, human-like intuition may develop within LLMs.

Furthermore, human intuition often results from grasping underlying statistical patterns and structures of the world in various respects, in an implicit way, from repeated experience^[8,25]. Likewise, repeated training of LLMs with a large amount of data facilitates implicit capturing within LLMs of statistical patterns and structures of the data^[26-27].

For example, Dasgupta et al^[28] and Saparov et al^[29]

showed that LLMs demonstrated content effects in reasoning similar to humans; Trott et al^[30] showed that LLMs could capture human intuition exhibited in false-belief experiments; and so on and so forth (see [6]).

Some may claim that there is a fundamental difference between intuition and language expressing intuition, or between intuition gained from the “real” world and whatever gained from linguistic texts. Such claims seem to unnecessarily divide language and the “real” world. To any individual, the world is made up of sensory-motor, linguistic, and other experiences, which are closely entangled and often mirror each other^[31-32]. Thus, one may gain intuition from any of these modalities. Furthermore, word meanings often emerge from use patterns^[26], so a “proximate” function such as next-token prediction in training LLMs gives rise to an “ultimate” function such as intuition^[27].

Now, beside intuition, what about another major type of implicit processes — instinct? Sun et al^[18] discussed how human motivation and consequent implicit action selection together capture instinct. It is mostly implicit and centers on the interaction of internally felt needs (motives) and external environmental factors in (implicitly) determining actions, thereby capturing much of characteristic behavioral patterns (e.g., personality). Beyond hardwired (innate) instincts, there are also acquired (learned) instincts, mostly through repeated experiences. What was argued earlier regarding intuition can be largely applied to instinct. In other words, instinct can develop within LLMs through repeated experiences with a large amount of training data.

However, can LLMs capture more than implicit processes (which include intuition and instinct)? LLMs do have some symbolic capabilities, including handling structures and compositionality to some extent (e.g., [29,31]). However, it is worth pointing out that LLMs (at least most of the currently existing ones) have only limited such capabilities. At the same time, human implicit mental processes have been shown to have likewise limited such capabilities^[8,33]. Therefore, it seems

unnecessary to posit anything more than implicit processes to explain limited symbolic capabilities of LLMs.

4.3 Many possibilities

For incorporating LLMs into Clarion, multiple possibilities exist, ranging from LLMs being at the periphery of Clarion to LLMs being at its core (cf. [20,34]):

(1) For dealing with perception: Multimodal LLMs handle implicit (early-stage) perception processes for Clarion. Given an input (e.g., a visual scene), LLMs generate linguistic and/or structured descriptions (in a form compatible with the internal representations of Clarion), for processing by Clarion.

(2) For dealing with language-based communication: In this case, natural language inputs are directed to LLMs, and LLMs (implicitly) process them to provide inputs to Clarion (in a form compatible with its internal representations). Clarion can then respond to them (in its internal forms), and LLMs generate natural language texts on that basis.

(3) For dealing with motor processes: LLMs (implicitly) generate motor commands, based on directives from Clarion (in its internal forms).

(4) For capturing memory: In this case, some modules of Clarion may be based on LLMs, which serve as (implicit) memory of various forms in Clarion (e.g., implicit semantic or procedural memory; [3]).

(5) For capturing all forms of implicit processes: LLMs can capture all implicit processes, ranging from intuition to instinct (as discussed before; minus motivation), aside from (implicit processes of) perception, motor control, and natural language communication (as mentioned above).

The last approach above evidently confers the most extensive role on LLMs within a cognitive architecture, as it essentially encompasses all the other possibilities mentioned. This approach is preferable due to its generality. This approach is also well justified theoretically (see Section 2 on dual-process theories and neuro-symbolic systems). Moreover, this approach is readily workable in any cognitive architecture that incorporates dual processes. See Fig. 2 for a sketch of

Clarion with LLMs incorporated, as will be explicated below.

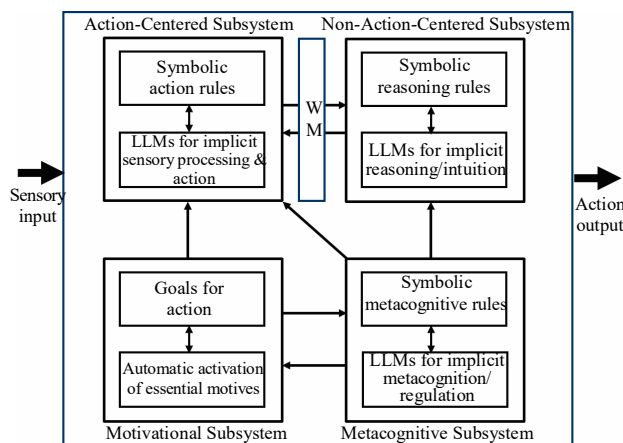


Fig. 2 The Clarion architecture incorporating LLMs

4.4 Capturing human-like dual processes

As discussed earlier, Clarion captures both implicit and explicit processes, as well as their interaction. Each type of process is not just tools; rather, each is a fundamental part of the human mind. Furthermore, explicit processes may guide implicit processes, but at the same time, they may also be guided by implicit processes. Thus, complex dynamic interaction occurs between these two types (cf. [7]).

The interaction processes, as elucidated in Clarion, can be mapped, when LLMs are incorporated into Clarion, to those between LLMs (capturing implicit processes) and explicit symbolic processes. That is, the Clarion dual processes, as discussed in Sections 2-3, can be leveraged for structuring the interaction between LLMs and symbolic processes.

With LLMs capturing implicit processes, top-down and bottom-up activation originally specified in Clarion become (1) prompts from symbolic processes at the top level to LLMs, and (2) outputs from LLMs (in a symbolic form) linked up with symbolic processes at the top level. As noted earlier, internal symbolic representations in Clarion are in the form of chunk nodes (each defined by a set of microfeatures; serving as a condition or a conclusion in explicit reasoning) or rules (with their conditions and conclusions specified by chunk nodes; acquired through learning). Using

embeddings from LLMs, which is an alternative, would require a much longer discussion and is thus omitted. Symbolic inputs to LLMs (e.g., from the top level) are prompts, which means prompt engineering and other techniques are needed. Note that LLMs can translate between natural language representations and internal representations of Clarion, which also facilitates the processing of external natural language inputs (e.g., queries, instructions, or situational descriptions) in both implicit and explicit ways.

In Clarion, with LLMs incorporated at the bottom level, explicit symbolic processes at the top level serve several roles: (1) providing prompts to the LLMs (at the bottom level) to initiate their processing; (2) verifying, rectifying, refining, or explaining the results from the LLMs (which may correspond to intuition or instinct), at the top level using explicit knowledge; (3) supplementing or even supplanting the results from the LLMs, using explicit knowledge at the top level, when situations call for it. Within Clarion, prompting, in a way, amounts to an inner dialogue within oneself^[35]. In this case, the dialogue mainly consists of repeated queries (from the top level) to one's inner self (e.g., intuition, at the bottom level). Through this dialogue, one's thoughts form, develop, and mature.

4.4.1 Reasoning in the NACS

Reasoning is a characteristically human capability. Dasgupta et al^[28], Saparov et al^[29], Trott et al^[30], and many others assessed reasoning by LLMs. Reasoning by LLMs tends to rely on statistical regularity or world knowledge, not strictly on logical rules. Although LLMs perform some reasoning, they are not yet capable of full human-level reasoning. When symbolic (explicit) and neural (implicit) processes co-exist in the NACS of Clarion, explicit symbolic processes add rigor to the system, through applying symbolic rules that perform more precise inferences in accordance with some logics or templates^[3-4]. Moreover, symbolic processes can invoke external tools more easily for performing certain specialized types of reasoning, such as mathematical derivation. When additional knowledge is needed in

reasoning, symbolic processes can also use external knowledge sources more easily. Results from symbolic processes may be used in conjunction with or instead of (supplanting) LLMs.

Various ideas in terms of prompt engineering also show the importance of explicit symbolic processes in guiding LLMs for reasoning. For example, Chain of Thought, Tree of Thought, or Algorithm of Thought prompts provide one-prompt guidance, from explicit (symbolic) to implicit (LLMs) processes. In addition, common search control strategies, such as depth-first, breadth-first, and A* search, can be implemented in Clarion from symbolic processes to LLMs (e.g., through multiple, iterative prompting). Furthermore, prompts from explicit symbolic processes to LLMs can involve even more complex, more sophisticated reasoning templates, which again are guidance to implicit processes. Xie et al^[36] showed that reasoning in LLMs could be enhanced through templates (given in a symbolic form) for performing analogical reasoning, problem decomposition, planning thinking, integrative thinking, and so on. In the original Clarion, there are a number of other reasoning methods (templates), such as forward chaining, backward chaining, constraint satisfaction, and so on, which can also be invoked by explicit symbolic processes (e.g., from the MCS; [3]). So, explicit symbolic processes can guide LLMs, in addition to the fact that they alone can perform explicit symbolic reasoning with precision, certainty, and other useful characteristics. Together, implicit-explicit interaction in Clarion leads to enhanced reasoning abilities, combining intuition and explicit processes.

4.4.2 Action in the ACS

Within Clarion, essentially the same kind of model (LLMs based on Transformers) can be used to capture a variety of implicit processes, including both intuition and instinct (as discussed before), in addition to various language, perception, and motor functions. Implicit processes for reasoning and for action are captured by separate (but interacting) LLMs, corresponding to the relation between the ACS and the NACS in Clarion as

described before^[3].

However, in Clarion, on top of implicit action selection processes (with LLMs), explicit symbolic processes are also present. Besides providing prompts to LLMs (e.g., as guidance), they may also perform explicit action selection with precision and other useful characteristics, in conjunction with or instead of LLMs.

In addition, planning of multi-step action sequences (an essential human capability) can also be carried out in Clarion through implicit-explicit interaction. For instance, Sun et al^[37] showed that explicit symbolic plans can emerge from neural learning. See also Park et al^[38] and other relevant work (details omitted due to space constraint).

4.5 Capturing human-like memory

There are a variety of memory in Clarion, as discussed in Section 3. LLMs, given their capabilities, serve well as implicit semantic memory and implicit procedural memory within Clarion (at the bottom levels of the NACS and the ACS, respectively), while the top levels of the NACS and the ACS serve as their explicit counterparts^[14].

However, episodic memory^[15] and working memory^[13] are often lacking in current LLMs. Clarion provides useful insights as to how to incorporate these needed memory systems, to lead to better overall performance^[34].

In Clarion, explicit episodic memory consists of an agent's experiences, in the form of *state*, *action*, *thought*, *payoff*, *next state*, and so on, along with a time stamp (indicating the time step). Here, *state* refers to what was perceived by the agent at the time step; *action* is what one did at that time step, including both external (e.g., verbal) and internal (e.g., mental) actions; *thought* refers to whatever happened in the NACS and the WM, including activated chunk nodes, associative rules that were applied, and so on; *payoff* is the reinforcement that one received after *action* was performed (based on the satisfaction of activated drives/motives within the MS); *next state* is what one perceived after *action* was performed. A retrieval function takes a partial

specification of state, action, and so on as the input and returns a subset of the memory that matches the input (to pass on to LLMs to generate linguistic descriptions, or to use as the basis for reasoning or learning).

In addition to episodic memory as an explicit list, neural networks are used for storing episodes in a more compact and more implicit way. Back-propagation neural networks have been used in the original Clarion, and LLMs are now used instead. A variety of techniques were developed within LLMs for condensing, summarizing, and integrating memory (e.g., [38]) that can lead to a condensed implicit episodic memory (which in turn forms a part of one's mental model). Such memory can be centered on Clarion internal representations or linguistic representations (or even embeddings).

Therefore, episodic memory in Clarion is dual-representational, mirroring the overall structure of Clarion. Each side serves a different function (specific episodes versus summarized experiences) and complements the other (as in the original Clarion).

Finally, working memory in Clarion is for storing information temporarily, facilitating the processing of the stored information (especially useful for the ACS), or for transferring information (especially between the ACS and the NACS, such as information from the ACS to be used as input to the NACS, or vice versa). The same working memory applies to Clarion with LLMs incorporated, to supplement the limited memory in LLMs and to transfer information between different subsystems (as in the original Clarion; [3]).

4.6 Capturing human-like motivation and regulation

In Clarion, as discussed earlier, the MS provides essential motives (drives) as basis for self-direction and self-regulation of behavior (which can then be carried out by the MCS). Work on Clarion shows that these motives (drives) are essential to accounting for a broad range of behavior, ranging from work performance to moral judgment (e.g., [39-40]). By being situated within Clarion, LLMs acquire human-like motivation. For

LLMs, this can enhance autonomy, behavioral consistency, value alignment with humans, mutual intelligibility with humans, and so on.

Furthermore, personality and emotion, on the basis of these motives as developed in Clarion (see [18]), lead to more human-like characteristics in LLMs. In Clarion, personality has been shown to be closely tied to human motivation, captured mostly based on motives/drives^[18]. Individual differences are explained largely by the differences in drive strengths. Such differences are consequently reflected in goals and actions. Implicit processes involved in this sequence roughly correspond to the notion of instinct (which can be captured by LLMs, as discussed earlier).

Likewise, emotion is also tied to motivation and action, as well as reasoning and metacognition. According to Clarion, emotion is generated as a function of both drive strengths and action potentials, with reasoning (e.g., for appraisal) and metacognition (e.g., for coping) as secondary processes (see [41]). Motivation is thus fundamental to emotion. Instinct and intuition (both of which are captured by LLMs) are an integral part of it.

Based on drives, goals, and other factors, the MCS monitors and regulates cognition and behavior^[19]. Metacognition and regulation can occur implicitly through LLMs (e.g., [38]) after initiation (prompting) by explicit symbolic processes. Meanwhile, explicit metacognition and regulation can also occur at the top level of the MCS. As a result of capturing metacognition in this way, Clarion can explain a much larger range of human behavior. By being situated within Clarion, LLMs acquire the capabilities.

Motivation, metacognition, self-direction, and self-regulation help to build a more complete model of human intelligence, as these are important elements of the human mind.

4.7 Capturing human-like learning

Learning has been touched upon earlier. Both top-down and bottom-up learning, as described in Section 3, are applicable to Clarion with LLMs. Along

with reinforcement learning and other forms of learning, they together lead to continuous, ongoing, incremental, real-time learning as in the original Clarion^[3].

A further opportunity offered by incorporating LLMs is that LLMs themselves may be prompted to generate useful symbolic rules (and other symbolic content) for the top level that can enhance the top level. This can be done right after pretraining of LLMs or during continuous learning (thus reflecting continuous learning). The generated symbolic content can contribute to explicit semantic or procedural memory (at the top level of the NACS or the ACS, respectively), supplementing symbolic knowledge obtained from externally provided instructions, external tools and sources, bottom-up learning, and so on^[3].

In addition, other forms of learning can also be applied to Clarion with LLMs. Learning from mistakes and learning from instructions can both be accomplished through episodic memory that feeds into LLMs (by following a memory-based approach; e.g., [36]). Curriculum learning (e.g., [42]) can be accomplished through the direction by the MCS in Clarion.

5 Concluding remarks

For computational models aiming to mimic or capture structures and processes of the human mind, hybrid neuro-symbolic models are crucial, inspired by (and at the same time substantiating) dual-process psychological theories. Their integration of neural networks and symbolic methods amplifies the strengths of both approaches while mitigating their respective weaknesses. A closer adherence to human psychology (based on mechanistic understanding of the human mind), as in computational cognitive architectures, may also allow such models to go beyond other existing approaches.

Neuro-symbolic, dual-process computational cognitive architectures incorporating LLMs can have capabilities beyond other cognitive architectures without LLMs, especially in their abilities to communicate (in

verbal or other forms) and in capturing a broader range of human intuition and instinct. As a result, for instance, cognitive architectures with LLMs can be used for dealing with real-world situations, as opposed to being limited to modeling small laboratory tasks as before.

In AI, there is also a good case for developing more human-like systems, better inspired and better guided by human psychology. Integrating cognitive architectures with LLMs is a promising approach for accomplishing this aim. This approach has the potential of leading up to human-level systems that include full symbolic capabilities, intrinsic motivation, human emotion, human personality, and other human characteristics (which are largely absent in current LLMs). This approach can also lead to better alignment with human values and a higher degree of autonomy. Apparently, reverse-engineering the human mind — the (presumably) best intelligent system around — has its advantages.

Acknowledgments

Thanks are due to Dr. Tingting Liu for her kind invitation to write this article. Thanks are also due to a number of colleagues for their useful discussions and suggestions.

References:

- [1] SUN R. The Cambridge handbook of computational psychology[M]. New York: Cambridge University Press, 2008.
- [2] SUN R. Duality of the mind[M]. Mahwah, NJ: Erlbaum, 2002.
- [3] SUN R. Anatomy of the mind[M]. Oxford, UK: Oxford University Press, 2016.
- [4] SUN R. Integrating rules and connectionism for robust commonsense reasoning[M]. New York: Wiley, 1994.
- [5] SUN R. Dual-process theories, cognitive architectures, and hybrid neural-symbolic models[EB/OL]. [2024-03-18]. <https://content.iospress.com/articles/neurosymbolic-artificial-intelligence/nai240720>.
- [6] SUN R. Roles of LLMs in the overall mental architecture [EB/OL]. [2024-10-26]. <https://arxiv.org/abs/2410.20037>.
- [7] EVANS J, FRANKISH K. In two minds: dual processes and beyond[M]. Oxford, UK: Oxford University Press, 2009.
- [8] REBER A S. Implicit learning and tacit knowledge[J]. Journal of Experimental Psychology: General, 1989, 118(3):219-235.
- [9] KAHNEMAN D. Thinking, fast and slow[M]. New York: Farrar, Straus and Giroux, 2011.
- [10] HÉLIE S, SUN R. Incubation, insight, and creative problem solving: a unified theory and a connectionist model[J]. Psychological Review, 2010, 117(3):994-1024.
- [11] SUN R, SLUSARZ P, TERRY C. The interaction of the explicit and the implicit in skill learning: a dual-process approach[J]. Psychological Review, 2005, 112(1):159-192.
- [12] PEW R W, MAVOR A S. Modeling human and organizational behavior: application to military simulations [M]. Washington, DC: National Academy Press, 1998.
- [13] BADDELEY A. Working memory[M]. New York: Oxford University Press, 1986.
- [14] SCHACTER D L. Implicit memory: history and current status[J]. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1987, 13(3):501-518.
- [15] TULVING E. Elements of episodic memory[M]. Oxford, UK: Clarendon Press, 1983.
- [16] MURRAY H A. Explorations in personality[M]. New York: Oxford University Press, 1938.
- [17] REISS S. Multifaceted nature of intrinsic motivation: the theory of 16 basic desires[J]. Review of General Psychology, 2004, 8(3):179-193.
- [18] SUN R, WILSON N. Roles of implicit processes: instinct, intuition, and personality[J]. Mind & Society, 2014, 13(1):109-134.
- [19] REDER L. Implicit memory and metacognition[M]. Mahwah, NJ: Erlbaum, 1996.
- [20] ROMERO O J, ZIMMERMAN J, STEINFELD A, et al. Synergistic integration of large language models and cognitive architectures for robust AI: an exploratory analysis[EB/OL]. [2024-08-26]. <http://arxiv.org/abs/2308.09830>.
- [21] COSMIDES L, TOOBY J. Beyond intuition and instinct blindness: toward an evolutionarily rigorous cognitive science[J]. Cognition, 1994, 50(1/2/3):41-77.
- [22] DREYFUS H L, DREYFUS S E. Mind over machine: the power of human intuition and expertise in the era of the computer[M]. Oxford, UK: Blackwell, 1986.
- [23] SUN R. Interpreting psychological notions: a dual-process computational theory[J]. Journal of Applied

- Research in Memory and Cognition, 2015, 4(3):191-196.
- [24] MCFARLAND D. Problems of animal behaviour[M]. Singapore: Longman, 1989.
- [25] HASHER L, ZACKS R T. Automatic and effortful processes in memory[J]. Journal of Experimental Psychology: General, 1979, 108(3):356-388.
- [26] DURT C, FROESE T, FUCHS T. Against AI understanding and sentience[EB/OL]. [2024-08-17]. <http://philsci-archive.pitt.edu/21983/>.
- [27] MOLLO D C, MILLIÈRE R. The vector grounding problem[EB/OL]. [2024-08-15]. <http://arxiv.org/abs/2304.01481>.
- [28] DASGUPTA I, LAMPINEN A K, CHAN S C Y, et al. Language models show human-like content effects on reasoning tasks[EB/OL]. [2024-09-26]. <http://arxiv.org/abs/2207.07051>.
- [29] SAPAROV A, HE H. Language models are greedy reasoners: a systematic formal analysis of chain-of-thought[EB/OL]. [2024-06-08]. <http://arxiv.org/abs/2210.01240>.
- [30] TROTT S, JONES C, CHANG T, et al. Do large language models know what humans know?[J]. Cognitive Science, 2023, 47(7):e13309.
- [31] PAVLICK E. Symbols and grounding in large language models[J]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2023, 381:20220041.
- [32] SØGAARD A. Understanding models understanding language[J]. Synthese, 2022, 200(6):443.
- [33] MACCHI L, BAGASSI M, VIALE R. Cognitive unconscious and human rationality[M]. Cambridge, MA: MIT Press, 2016.
- [34] SUMERS T R, YAO S Y, NARASIMHAN K, et al. Cognitive architectures for language agents[EB/OL]. [2024-07-13]. <http://arxiv.org/abs/2309.02427>.
- [35] VYGOTSKY L S. Thought and language[M]. Cambridge, MA: MIT Press, 1962.
- [36] XIE Y Z, XIE T, LIN M X, et al. OlaGPT: empowering LLMs with human-like problem-solving abilities[EB/OL]. [2024-07-21]. <http://arxiv.org/abs/2305.16334>.
- [37] SUN R, SESSIONS C. Learning plans without *a priori* knowledge[J]. Adaptive Behavior, 2000, 8(3/4):225-253.
- [38] PARK J S, O'BRIEN J C, CAI C J, et al. Generative agents: interactive simulacra of human behavior[EB/OL]. [2024-09-03]. <http://arxiv.org/abs/2304.03442>.
- [39] BRETZ S, SUN R. Two models of moral judgment[J]. Cognitive Science, 2018, 42(Suppl. 1):4-37.
- [40] SUN R, BUGROV S, DAI D. A unified framework for interpreting a range of motivation-performance phenomena [J]. Cognitive Systems Research, 2022, 71:24-40.
- [41] SUN R, WILSON N, LYNCH M. Emotion: a unified mechanistic interpretation from a cognitive architecture [J]. Cognitive Computation, 2016, 8(1):1-14.
- [42] ELMAN J L. Learning and development in neural networks: the importance of starting small[J]. Cognition, 1993, 48(1):71-99.

计算认知架构、双过程理论和大语言模型

SUN Ron

(伦斯勒理工学院 认知科学系, 美国 纽约 特洛伊 12180)

摘要: 介绍了一个认知科学和计算心理学领域里用于人类思维综合计算理论的计算认知架构. 这一名为 Clarion 的认知架构得到了心理学数据、发现和基本理论的验证. 它的一个重要理论背景是双过程理论, 这也是该模型以混合神经符号方式采用整体两级结构的原因. 此外, 鉴于人工智能和计算技术的最新进展, LLM 被纳入模型中, 以更好捕捉人类的直觉和本能(以及一般意义上的隐性过程), 并进一步增强 Clarion. 整合 Clarion 和 LLM 还可以帮助开发更强大、更可靠、更像人类的人工智能系统. 总的来说, 倡导采用多学科方法来开发更好的认知科学和人工智能模型.

关键词: 认知架构; 大语言模型; 双过程; 心理学; 人工智能

(责任编辑 韩 超)