An Architectural Approach to Metacognition

Christian Lebiere, Robert Thomson, Andrea Stocco, Mark Orr, Donald Morrison

1 Introduction

1.1 Classes of Metacognitive Processes

Confusingly, researchers have used the term "metacognition" to indicate a number of heterogeneous mental phenomena. At the highest possible level, the label encompasses any kind of knowledge and mental processes that have one's own cognition as its object [16].

This definition includes low-level evaluations of one's own memory (feeling or knowing, familiarity, and meta-memory judgements) and performance (feeling of confidence in one's own decisions) as well as complex, deliberate reflections about one's own cognition (such as assessing whether you would be able to pass a test), scaling up to complex forms of reasoning (e.g., solving an induction puzzle, such as the "three wise men", which requires considering different scenarios in which the thinker might have different amounts of knowledge).

1.2 Automatic vs. Deliberate Forms of Metacognition

To impose some order onto these phenomena, several authors have proposed their own taxonomies of metacognitive processes [16, 44, 65]. For convenience, in this chapter we will draw a line between meta-cognitive processes that are *deliberate* and meta-cognitive processes that are *automatic*. In the terms popularized by Daniel Kahneman, deliberate metacognitive processes are examples of "System 2" processing, while automatic metacognitive phenomena are examples of "System 1" processing. Within the metacognition literature, this taxonomy is perhaps closest to Flavell's distinction between "metacognitive experience" and "metacognitive knowledge"; however, as the remainder of the chapter will make it clear, the difference between automatic and deliberate metacognition is more directly related to a computational approach.

Most of the deliberate forms of meta-cognition can be thought of as a form of reasoning: When agents are engaged in this form of metacognition, they are effectively "thinking about thinking". They might, for example, search their memory to assess whether they possess some particular knowledge or skill, or whether they can explain how they performed a task. In a subset of paradigms that are often singled out as quintessentially meta-cognitive, the nature of reasoning is *recursive*: When an agent is engaged in a meta-cognitive task, they are internally simulating what they would do in a specific situation. A specific variant of these tasks involves reasoning not only about one's own mind, but about other minds as well. This is the case of collaborative or competitive tasks such as the prisoner's dilemma, in which an agent might engage in thinking about what they would do in response to another agent's actions. This particularly elaborative form of reasoning, which involves taking both one's perspective and that of a different agent, is also known as *Theory of Mind* [34, 59].

The types of metacognitive processes that are automatic, on the other hand, do not involve any form of reasoning. They involve, instead, the perception of specific signals that mark the status of cognitive processes, and which are likely generated spontaneously and automatically.

One such example is the phenomenon known as feeling of knowing [29]. When asked a question, individuals can often respond whether they know the answer faster than they can provide the answer itself. An extreme example of this feeling of knowing is the frustration associated with the "tip of the tongue" effect [8], i.e. the familiar feeling of almost remembering a particular fact (e.g. the name of an actor in a movie) without being actually able to retrieve it. A second example is the feeling of response conflict, or the feeling of mental impasse that occurs when different responses are competing for execution, or when a prepotent one needs to be suppressed. Note that these forms of meta-cognition are automatic in the sense that they are signals generated without the agent's intention. Their use might still be deliberate, and their perception might spurn higher-level forms of thinking.

One striking distinction between these two classes of metacognitive phenomena is their different speed: deliberate metacognitive processes are slow, while automatic ones are fast. Consider, for example, the case of meta-cognitive assessment involved with asking someone whether they know the answer to a fact. This typically involves the deliberate scanning of one's own memory and some time to respond. In contrast, the feeling of knowing the effect, which functionally contains the same information (that is, whether a memory is present) is fast and automatic, to the point that individuals can correctly state they know an answer before they have actually retrieved it [64].

Another relevant distinction is that automatic processes have well known, idiosyncratic, and localized neural signatures [49, 75]. In contrast, virtually all the types of reasoning discussed before share similar neural signatures, typically involving the "multi domain system"–a network of interconnected brain regions that is involved in virtually all forms of difficult mental processing, and includes regions involved in attention and working memory [13].

1.3 Implications for Cognitive Architectures

From a cognitive architecture perspective, the distinction between automatic and deliberate processes is particularly important. In cognitive architectures, the most automatic mechanisms (for example, those that control the execution of procedural knowledge or the access to declarative memory) can be considered as architectural primitives and fundamental features of the system, while the least automatic and more deliberate processes can usually be simulated using the architecture's built-in primitives [69]. Thus, from an architectural perspective, there is no reason to believe that any deliberate, System 2-like form of metacognition is, in itself, any different from any other form of reasoning. But the level of automaticity that is characteristic of System 1-level processes, together with their specific neuronal signals, suggests that these type of metacognitive signals are themselves part of the fundamental functions of the architecture.

Of course, the distinction we have outlined is not clear-cut. Automaticity *per se* does not require the existence of any primitive; reading, for example, is a highly automated capacity (it is almost impossible *not* to read a word in front of us, which is the reason why the Stroop effect exists: MacLeod 40) and yet the cognitive system is not born with a "reading" module—it takes years of practice to learn how to read, and the process results in substantial rewiring of large portions of the brain.

Conversely, it might be argued that at least some forms of deliberate, System-2-level metacognitive phenomena do indeed count as architectural primitives. Consider, for example, the set of reasoning tasks that are described as Theory of Mind. A few paragraphs above, we presented them as some of the higher and most sophisticated forms of reasoning; in fact, younger children and animals consistently fail at them. And yet, many researchers have pointed out that Theory of Mind tasks might be rooted, at the neural level, in the existence of mirror neurons . Mirror neurons in the primate brain have the unique property of firing both when an animal is executing a specific movement and when the animal sees the same movement executed by a different agent [55]. Because of their unique property, they have been speculated as the foundational mechanisms by

which primates, and perhaps other animals, understand other agents' intentions and take other agents's perspective [23]. Mirror neurons are definitely a basic, "architectural" property of the human brain; thus, if they are necessary for TOM tasks, one must conclude even higher-level reasoning tasks might ultimately be rooted in some fundamental metacognitive mechanism in the human brain.

1.4 Metacognitive Measures

The author Alfred North Whitehead summarized metacognitive awareness best: "[t]he purpose of thinking is to let the ideas die instead of us." [1] Metacognition provides an imperfect set of internal feedback and calibration signals that help provide robustness to decision-making in the face of resource-limited cognitive capabilities in complex, open-ended environments. Two factors greatly influence this robustness: 1) the relative confidence in our (or others') capability to successfully complete a task, and 2) an internal estimate of the effort required to complete a task. Perhaps the most well-studied measure of metacognition is the confidence judgment. For the purpose of this paper, we will not describe potential differences between confidence and certainty (see 52). A common feature of confidence judgments is that there is an initial tendency for novice participants to exhibit overconfidence when learning a new task (60; also see 70, 62), followed by a period of underconfidence with practice when the task is challenging [33, 27], and continued overconfidence in relatively-easier tasks [58]. Furthermore, when participants know that they must generate a confidence score, their response times tend to slow and their responses tend to get more accurate [60]. This means that explicitly generating a confidence judgment actually alters the decision-making process.

What makes confidence so interesting is that it can be dissociated from accuracy. Confidence leak occurs when the confidence judgment from recent trials intrudes on current confidence judgments, whereas the same is not necessarily the case for accuracy. Similarly, cue salience can differentially impact confidence and accuracy in the same task [66]. In perceptual tasks, a positive evidence bias occurs when confidence over-weights the objective evidence whereas task accuracy is unimpacted (effectively a confirmation bias which occurs for confidence judgments as opposed to accuracy).

A second measure of metacognition is the ability to estimate the effort required to complete a task. Although similar to confidence, this ability is not perfect. This prediction influences how much effort we put towards studying for a test [17, 32] and our relative attention allocation in perception [57]. Similar to

[1] Astute readers using their metacognitive skills will recognize that this is a frequent, but apocryphal, attribution to Whitehead, and is probably a paraphrase of a passage in [51]

confidence, novice participants tend to overestimate their own capabilities and underestimate time demands when assessing the effort required to complete a novel task. Recent evidence shows that ready access to the internet may exacerbate this mis-estimation as the availability of information online provides a false sense of knowing [14, 2].

1.5 Functional Benefits

What sets metacognitive capabilities apart is not only that we are able to identify our own (or other's) internal states, but that we are able to evaluate and reason over these states [17, 50]. This provides a robustness that is unique to human decision-making (compared with AI-based methods). One's internal awareness of confidence/certainty works with an estimate of effort to integrate with simulation and experience to form a prediction error feedback signal to drive (often unsupervised) learning [53, 12, 74]. Specifically, metacognition mediates the relationship between executive functioning and self-regulated learning (18; for a review see 32). The major benefits of this is that we are able to allocate resources based on specific goals, and failures of this signal (i.e., mis-calibration) provides motivation to adapt by increasing skill, increasing effort, and/or seeking external resources (e.g., social interactions). It also provides for an estimate of whether the cost of the goal is worthwhile ('is the juice worth the squeeze?').

What specific benefits does this provide? First off, metacognitive signals provide the motivation to change social attitudes, update beliefs, and enhance strategic learning outcomes to match our current goals [50, 67, 7]. Specifically, it provides an estimate to balance effort with reward to maximize a goal, which can also trigger when to learn (e.g., how hard to study; 9). With resource constraints, it provides the drive to satisfice (see 61 for a discussion). [1] showed that our tendency to underestimate the time requirements when studying complex topics is due not only to estimating confidence of how close one was towards achieving a predetermined aptitude (i.e., a judgment of learning), but also a top-down assessment of how much time it should have taken with reduced motivation and premature completion of learning. While possibly seen as a negative, this actually avoids the sunk-cost fallacy when one's judgment of required cognitive effort is mis-calibrated.

Another benefit is that we can use metacognitive estimates to distinguish between things that we do not know and things that we have forgotten [26, 24]. If a given event should have been memorable and was not recalled, then it was likely something that one had never known. Conversely, if a given event should

not have been memorable and was not recalled, it was more likely to be judged to be forgotten.

While the current section has focused on several functional benefits that metacognitive awareness provides, limits of our ability to use metacognitive signals can sometimes lead us to validate negative perceptions and use bias-prone heuristics [43]. Metacognitive signals provide motivation to change attitudes/beliefs and can enhance learning, but can just as easily cause us to harden our viewpoints when biases get in the way. Blindness to our biases and lack of skills (e.g., the Dunning-Kruger effect; 15) leads to poor calibration, and poor performance on self-regulation.

2 Metacognition in Cognitive Architectures

A major challenge in defining metacognition is to distinguish it from "regular" cognition. Metacognition has been defined as "the ability to monitor and adaptively control one's cognitive processing or thinking about thinking" [48]. However, human cognition has many ways of monitoring and adapting its behavior in various contexts, many of which are viewed as part of everyday behavior. What is needed to properly define metacognition is therefore a reference baseline for the structures and mechanisms that constitute "regular" cognition. Cognitive architectures provide a natural integrative framework for that definition.

Following his insight that a divide-and-conquer approach to modeling human cognition could not provide a road map to its ultimate goal [45], Newell proposed the concept of unified theories of cognition, implemented computationally as cognitive architectures [46]. The last five decades has seen a number of diverse proposals to explore the luxuriant space of cognitive architectures [30]. We will focus here on a particular architecture, ACT-R, specifically aimed at modeling human cognition from a neuro-psychological perspective [5].

The ACT-R cognitive architecture (Figure 0.2) is composed of a set of modules with dedicated functionality, localized in specific brain areas [4]. The central module, procedural memory, controls the flow of information between other modules using condition-action pairs, aka production rules. The interface between modules consists of a set of buffers, collectively known as working memory. Each buffer is attached to a particular module and can hold one piece of information, known as a chunk, at a time. When a production rule is selected, it can change the content of one or more buffers, triggering corresponding processes in the associated module. Example actions include shifting attention or recognizing an object in the visual module, or retrieving a piece of information

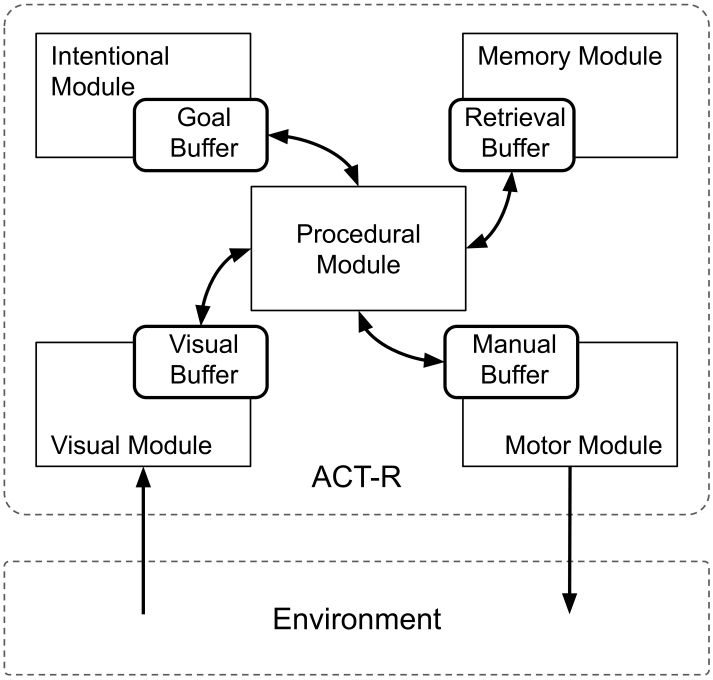

Figure 0.2 ACT-R Cognitive Architecture

from the long-term declarative memory module. Those processes typically operate on a combination of symbolic knowledge structures, such as the chunks in memory that are represented as sets of attribute-value pairs, with attached subsymbolic quantities that are learned statistically to reflect the structure of the environment and control access to information. The results of those processes ultimately become available in the corresponding buffer. The procedural module can then detect the asynchronous change and attempt to match the condition part of its production rules against the new state of working memory, and the cognitive cycle repeats.

A common assumption is that awareness is associated with information present in the buffers while the internal contents of the modules are not explicitly accessible. For instance, while we can retrieve information from long-term memory, that process is probabilistic and approximate and we cannot directly know the contents of our memory. In particular, we do not have explicit access to the subsymbolic quantities (e.g., the activation of memory chunks that

determine their probability and latency of retrieval) that modulate cognitive processes. However, that absolute encapsulation of the contents of cognitive modules is difficult to reconcile with the limited degree of awareness of our own cognitive processes described in the previous section. Thus, we propose the following conjectures defining the nature of metacognition in the context of cognitive architectures:

Conjecture 1: Metacognition involves extracting information about module processes

It is essential to emphasize the fundamental distinction between information about processes as opposed to information about knowledge, even though of course processes operate on knowledge. Information about knowledge, e.g. the activation of a particular chunk, is local, specific to that specific item, and encapsulated in the module. Information about a process reflects the entirety (or at least a wide range) of the module content, e.g., the activation of all the chunks competing in a retrieval request, and can be reflected in a metacognitive signal.

Conjecture 2: Metacognitive information is quantitative and approximate rather than symbolic in nature

Metacognitive information about processes is not symbolic but rather expresses graded quantities that provide functional insight into those processes such as probability of success, degree of competition, confidence in the answer, and salience of various relevant factors. Despite the quantitative nature of the information, it is not subsymbolic but rather cognitively accessible and actionable.

Conjecture 3: Metacognitive information is available in working memory for cognitive processing

Even though metacognition is a distinct aspect of human cognition, it continues the architectural pattern of making information extracted from cognitive modules available in working memory for further processing. One possibility is to extend the current distinct distinction between content buffers that hold the result of module processes and state buffers that currently hold the state of the module, i.e. whether the module is currently free or busy (to prevent requesting another operation when a module is currently busy). Note that these conjectures are about the monitoring role of metacognition. No additional assumption is

made about any additional metacognitive processes that would act on the result of that monitoring:

Conjecture 4: General cognitive processes are sufficient to respond to a situation detected by metacognitive monitoring

Once metacognitive signals are made available into working memory, the information they convey can be processed using the standard mechanisms that an architecture can employ to process any other information. In particular, all the forms of deliberate metacognition (including inference about one's knowledge and theory of mind) can be implemented using the standard tools that allow any cognitive architecture to reason and problem-solve.

3 Computational Cognitive Models of Metacognition

We describe here some examples of the kind of metacognitive ability conjectured in the previous section. While those models were developed in the context of the ACT-R cognitive architecture, they were prototypes that were not integrated in the specific way conjectured above. Rather, these examples are intended to illustrate the kind of metacognitive signals that can be extracted within the cognitive architecture framework and the functional benefits that they could confer.

3.1 Visual Perception

One of the key functions of visual perception is object categorization. The ACT-R visual module operates in a way reflecting the principles of human vision: it directs attention through a request in the visual-location buffer then requests the object at that location in the visual field to be encoded with the result to be made available in the visual buffer. Object recognition in the ACT-R visual module is fairly primitive and only operates on symbolic representations typical of a computer screen used in cognitive psychology experiments. While in that context object recognition can operate according to a well-defined standard, the same is not true in the real world, where interpreting the content of visual scenes is rife with uncertainty. In that context, obtaining estimates of the quality of the output of the recognition process, such as probability of the accuracy of the judgment, would be highly desirable. However, those estimates would be based on an underlying model whose (in)correctness would likely be correlated

with that of the response itself. Instead, indirect measures of correctness such as confidence might be more realistic.

To explore those possibilities, we combined the ACT-R cognitive architecture with a neurally plausible vision model reflecting the structure of the human visual cortex [28]. That model, based on the Leabra neural architecture, is a massively parallel neural network similar to convolutional neural networks but using a neurally plausible algorithm combining top-down error correction with bottom up Hebbian self-organization [47]. The original approach to this integration [72] is to extract global activation quantities from the neural network, including the average activation of the last layer featuring a distributed representation (called the IT layer by analogy to its equivalent in the human visual system) and the maximum activation of the final winner-take-all layer. Those metacognitive quantities are used to build a classifier to detect low-confidence categorizations. When categorization confidence falls below a threshold, a new view of the object is taken and a new cycle of categorization is initiated. The iterative process continues until the confidence threshold is reached and a final categorization is accepted, or a cycle limit is exceeded and the system gives up. The second approach [73], is similar to the first but, instead of extracting scalar metacognitive signals of limited resolution, it introspects into the IT layer to extract the most abstract distributed representation of the object used for categorization. That representation is then associated with the object identification in the declarative memory of the cognitive architecture. The cognitive model can then use memory retrieval processes to determine if the representation is similar enough to other objects of the same type. If that similarity judgment falls below a threshold, then it is determined that the object belongs to a new category. The cognitive model then provides top down input to the visual model to create a new object category and start training it with the current object. This combined system proves capable of accurately learning entirely new object categories without any external supervisory feedback.

Two aspects of this approach to visual metacognition are worth emphasizing. The first is that a hybrid symbolic-neural system provides a natural implementation for the kind of perspective that we are proposing, with the metacognitive monitoring signal extracted from the internal dynamics of the neural sub-system and made available to the symbolic cognitive system. The second noteworthy aspect is that the intervention following the generation of the metacognitive signal detecting an exception condition (a failure to properly recognize an object due an insufficient view or lack of training on that specific category, respectively) does not require additional metacognitive capabilities. Instead, it uses standard cognitive mechanisms such as declarative memory representation and

pattern matching to implement strategies designed to remediate the situation (taking another perspective, or learning a new object category, respectively).

3.2 Declarative Memory

Declarative memory is the long-term repository of experience and knowledge in the cognitive architecture. ACT-R does not draw a distinction between episodic memory, holding first-person experiences, and semantic memory, holding more abstract forms of knowledge. But, like other modules, it shares the bottleneck of that potentially vast amount of information being accessed through a limited-capacity buffer that can only hold a single chunk at a time. One can think of it as a form of focusing similar to visual attention, where a pattern is placed in the retrieval buffer requesting the most relevant chunk to be retrieved and deposited back in the same buffer. Usually, the pattern requesting the retrieval is underspecified, leaving many potential chunks in declarative memory eligible to be retrieved. As mentioned previously, the eligible chunk with the highest activation is retrieved, reflecting several statistical factors such as frequency and recency of access, associative priming from the current working memory context, and degree of match to the requested pattern. The latency of declarative memory retrieval is typically on the order of hundreds of milliseconds, meaning that only a few chunks of long-term information can be retrieved per second. Providing relevant information given these constraints is an incredibly difficult problem to solve for the memory system. Flawless performance is far from assured, which makes the metacognitive ability to introspect into declarative memory retrieval processes highly desirable in order to inform cognitive strategies designed to improve the performance and robustness of the system. For instance, in the case of the feeling of knowing, a measure of the probability or closeness of success in the event of retrieval failure could trigger future attempts at retrieval, perhaps combined with a strategy such as priming from related information that could improve the probability of a successful retrieval.

We illustrate the kind of metacognitive signals that can be extracted from declarative memory retrieval processes and the use that can be made of them using an example of decision making applied to cyber security [39]. The cognitive model follows an approach called Instance-Based Learning (IBL; [25]) which makes decisions by generalizing from instances of experiences held in declarative memory using memory retrieval processes, in particular a mechanism called blending [38] that aggregates multiple chunks to produce a probabilistic expected outcome. The task is a simulated cyber security experiment involving an insider attack on a number of potential targets with distinct characteristics. After a target is chosen, a deceptive signal is given to try to convince the intruder

not to attack. The original deceptive signal generated using a Stackelberg game theory paradigm was found to be suboptimal against actual human subjects. Instead a personalized cognitive model aligned against a specific attacker's behavior trace is shown to be a better predictor of individual human behavior. The model is then used to optimize the deceptive signal to a specific attacker by extracting from the model a metacognitive signal of the strength of belief (i.e., trust) in the signal. That level of strength is computed from the activation of the relevant beliefs in memory, specifically past instances of success and failure in attacking in the presence or absence of a signal, which reflects the frequency and recency of past experiences. That metacognitive quantity does not provide direct access to the activation levels of specific memories, which would be cognitively implausible, but rather the relative strength of one set of experiences against another.

A second metacognitive signal called cognitive salience is extracted from the model to quantify the relative reliance of the various target characteristics in the selection process. It is computed as the derivative of the output of the blending process used to generate expected reward for attacking each target with respect to the various features defining those targets. Again, it does not provide access to any specific subsymbolic quantity in memory but rather reflects the overall state of memories relevant to the retrieval process. This concept of cognitive salience was originally defined for the purposes of explainable AI to explain to a human user the underlying basis of decisions made by an intelligent agent [42]. In this case, the cognitive salience values can be used to shift coverage toward targets whose characteristics are more salient to a specific user. As for the perceptual modules described in the previous section, the interventions driven by signals such as trust or salience extracted from memory retrieval processes are not metacognitive in nature but rather can be enacted by the same kind of cognitive decision strategies that are the object of the introspection.

4 Discussion

In this chapter, we have argued that metacognitive phenomena can be divided into automatic and deliberate, and and that, from a cognitive architecture point of view, the first ones can be conceived as fundamental architectural primitives, while the second ones can be achieved using the architecture's standard processes, once signals of the first type are detected. Although the conjectures and the theory of metacognition exposed in this paper were formulated in reference to cognitive architectures, they can be extended at lower and higher levels of analysis.

At the lower level, our conjectures are broadly compatible with the principles of predictive coding in neural systems [54, 21], that is, the idea that the brain is organized to maximize homeostasis (or, equivalently, to minimize its "free energy": Friston 20) by minimizing surprisal and maximizing the predictability of the surrounding environment. Violations of expectations can be used by an agent to modify its behavior—for instance, by changing its own decision policy or moving to a new environment [22]

Within this framework, it is possible to conceive of the specific metacognitive signals that we have reviewed as violations of expectations about how well the cognitive system itself is working. For example, the feeling of knowing would be a violation of expectations about retrieving a memory; the "Aha!" experience in problem solving would correspond to a sudden change in the expectation of solving a problem; and the sense of confidence or uncertainty about a decision would corresponding to a sudden change in the perceived effectiveness of a decision.

One obstacle to reconciling this hypothesis with a cognitive architecture viewpoint is that predictive coding, as its name implies, requires the cognitive system to be continually making predictions, and cognitive architectures are not explicitly built upon this principle. That does not mean, however, that cognitive architectures do not make predictions: in fact, many aspects of a cognitive architecture can be seen as implicit predictions about future states of the world. In ACT-R, for example, each declarative memory has an associated scalar variable called activation, which represents the log odds of this memory being needed at that particular moment in time [3]. The distribution of activations across all memories, therefore, represents an implicit prediction about what should be expected in the environment [6]. Similarly, each procedural memory has an associated scalar quantity, its utility, that is computed through reinforcement learning and computes the expected future reward of the corresponding skill. Again, such a term implicitly defines a prediction, and is temporally adjusted by reducing the mismatch between predicted and actual rewards [63]. Thus, it is possible to relate our architectural view of metacognition to the larger framework of predictive coding and active inference.

At a higher level, a wide range of capabilities have been integrated into cognitive architectures over the last 50 years of exploration of the design space [30]. A number of architectures incorporate metacognitive features such as monitoring of internal resources and extracting confidence values, e.g., CLARION [68], Companions [19], and Soar [35]. Many architectures also include other features that are often associated with metacognition but that we do not consider here to be inherently metacognitive, such as temporal representation of alternative solutions, changing task priorities, storing and using traces of ex-

ecution, improving analogy or problem solving, and general aspects of theory of mind.

A complementary direction of research is to apply the insights into metacognition achieved in the context of cognitive architectures to Artificial Intelligence systems that share many commonalities despite apparent differences. As previously mentioned, cognitive salience is a technique based on blended retrievals which can be used to extract which features contribute to a given decision. When applied to AI algorithms via model tracing (having a cognitive model make the same decisions as an AI, or human, agent), it is possible to investigate the relative contribution of each feature to the given decision, in a manner somewhat analogous to SHAP values [41]. A preliminary investigation of using cognitive salience in the context of models of intrusion detection in a network defense application has shown potential as a metacognitive signal to understand feature importance [71].

A recent convergence in the broad diversity of cognitive architectures has recently prompted an attempt at formalizing an emerging consensus called the Common Model of Cognition (CMC; [36]). A working group organized in the context of that effort summarized various aspects of metacognition [31] but, because of the relative lack of maturity and absence of consensus in approaches to metacognition, it has not been included in current proposals. However, recent efforts to elaborate and integrate a theory of emotions with the CMC have led to a proposal for a treatment of metacognition focused on appraisal theory that is generally compatible with the approach of the current chapter [56, 37] and builds off some earlier research integrating affective components to cognitive models via modeling the physiological substrate of cognition [10, 11]. Integrating physiological modeling into congitive models provides a mechanism to model affective features such as emotion in addition to factors such as fatigue and stress. These enterprises further reinforce the belief that metacognition is a highly active area of research that can be integrated within the existing framework of cognitive architectures and is likely to bear many fruits in coming years.

5 Acknowledgements

This research was sponsored by the Department of Defense Basic Research Office award HQ00342110002, the Army Research Office MURI grant Number W911NF-17-1-0370, and the Office of Naval Research MURI Award N0001422MP00465. The views expressed in this work are those of the authors and do not necessarily reflect the official policy or position of the United

States Military Academy, Department of the Army, Office of Naval Research, Department of Defense, or U.S. Government.

References

[1] Ackerman, Rakefet. 2014. The diminishing criterion model for metacognitive regulation of time investment. *Journal of experimental psychology: General*, **143**(3), 1349.

[2] Ackerman, Rakefet, and Goldsmith, Morris. 2011. Metacognitive regulation of text learning: on screen versus on paper. *Journal of experimental psychology: Applied*, **17**(1), 18.

[3] Anderson, John R. 1990. *The adaptive character of thought*. Lawrence Erlbaum Associates.

[4] Anderson, John R. 2007. *How can the human mind occur in the physical universe?* Oxford University Press, USA.

[5] Anderson, John R., and Lebiere, Christian. 1998. *The Atomic Components of Thought*. Mahwah, N.J.: Lawrence Erlbaum Associates.

[6] Anderson, John R, and Schooler, Lael J. 1991. Reflections of the environment in memory. *Psychological science*, **2**(6), 396–408.

[7] Biggs, John. 1988. The role of metacognition in enhancing learning. *Australian Journal of education*, **32**(2), 127–138.

[8] Brown, Alan S. 1991. A review of the tip-of-the-tongue experience. *Psychological bulletin*, **109**(2), 204.

[9] Colombo, Barbara, Iannello, Paola, and Antonietti, Alessandro. 2010. Metacognitive knowledge of decision-making: An explorative study. *Trends and prospects in metacognition research*, 445–472.

[10] Dancy, Christopher L. 2013. ACT-RΦ: A cognitive architecture with physiology and affect. *Biologically Inspired Cognitive Architectures*, **6**, 40–45.

[11] Dancy, Christopher L, Ritter, Frank E, Berry, Keith A, and Klein, Laura C. 2015. Using a cognitive architecture with a physiological substrate to represent effects of a psychological stressor on cognition. *Computational and Mathematical Organization Theory*, **21**, 90–114.

[12] Desender, Kobe, Boldt, Annika, and Yeung, Nick. 2018. Subjective confidence predicts information seeking in decision making. *Psychological science*, **29**(5), 761–778.

[13] Duncan, John. 2010. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, **14**(4), 172–179.

[14] Dunn, Timothy L, Gaspar, Connor, McLean, Daev, Koehler, Derek J, and Risko, Evan F. 2021. Distributed metacognition: Increased bias and deficits in metacognitive sensitivity when retrieving information from the internet.

[15] Dunning, David. 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. Pages 247–296 of: *Advances in experimental social psychology*, vol. 44. Elsevier.

[16] Flavell, John H. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, **34**(10), 906.

[17] Fleming, Stephen M. 2024. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, **75**, 241–268.

[18] Follmer, D Jake, and Sperling, Rayne A. 2016. The mediating role of metacognition in the relationship between executive function and self-regulated learning. *British Journal of Educational Psychology*, **86**(4), 559–575.

[19] Forbus, Kenneth D, and Hinrichs, Thomas R. 2006. Companion cognitive systems: a step toward Human-Level AI. *AI magazine*, **27**(2), 83–83.

[20] Friston, Karl. 2010. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, **11**(2), 127–138.

[21] Friston, Karl, and Kiebel, Stefan. 2009. Predictive coding under the free-energy principle. *Philosophical transactions of the Royal Society B: Biological sciences*, **364**(1521), 1211–1221.

[22] Friston, Karl, FitzGerald, Thomas, Rigoli, Francesco, Schwartenbeck, Philipp, Pezzulo, Giovanni, et al. 2016. Active inference and learning. *Neuroscience & Biobehavioral Reviews*, **68**, 862–879.

[23] Gallese, Vittorio, and Goldman, Alvin. 1998. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, **2**(12), 493–501.

[24] Ghetti, Simona. 2003. Memory for nonoccurrences: The role of metacognition. *Journal of Memory and Language*, **48**(4), 722–739.

[25] Gonzalez, C., Lerch, J. F., and Lebiere, C. 2003. Instance-based learning in dynamic decision making. *Cognitive Science*, **27**, 591–635.

[26] Grimaldi, Piercesare, Lau, Hakwan, and Basso, Michele A. 2015. There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making. *Neuroscience & Biobehavioral Reviews*, **55**, 88–97.

[27] Hanczakowski, Maciej, Zawadzka, Katarzyna, Pasek, Tomasz, and Higham, Philip A. 2013. Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language*, **69**(3), 429–444.

[28] Jilk, David J, Lebiere, Christian, O'Reilly, Randall C, and Anderson, John R. 2008. SAL: An explicitly pluralistic cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, **20**(3), 197–218.

[29] Koriat, Asher. 2000. The feeling of knowing: Some metatheoretical implications for consciousness and control. *Consciousness and cognition*, **9**(2), 149–171.

[30] Kotseruba, Iuliia, and Tsotsos, John K. 2020. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, **53**(1), 17–94.

[31] Kralik, Jerald D, Lee, Jee Hang, Rosenbloom, Paul S, Jackson Jr, Philip C, Epstein, Susan L, Romero, Oscar J, Sanz, Ricardo, Larue, Othalia, Schmidtke, Hedda R, Lee, Sang Wan, et al. 2018. Metacognition for a common model of cognition. *Procedia computer science*, **145**, 730–739.

[32] Krieger, Florian, Azevedo, Roger, Graesser, Arthur C, and Greiff, Samuel. 2022. Introduction to the special issue: the role of metacognition in complex skills-spotlights on problem solving, collaboration, and self-regulated learning. *Metacognition and Learning*, **17**(3), 683–690.

[33] Kubik, Veit, Jemstedt, Andreas, Eshratabadi, Hassan Mahjub, Schwartz, Bennett L, and Jönsson, Fredrik U. 2022. The underconfidence-with-practice effect in action memory: The contribution of retrieval practice to metacognitive monitoring. *Metacognition and Learning*, **17**(2), 375–398.

[34] Kuhn, Deanna. 2000. Theory of mind, metacognition, and reasoning: A life-span perspective. Pages 301–326 of: Mitchell, P, and Riggs, K. J. (eds), *Children's Reasoning and the Mind*. East Sussex Psychology Press Ltd.

[35] Laird, John E. 2019. *The Soar cognitive architecture*. MIT press.

[36] Laird, John E, Lebiere, Christian, and Rosenbloom, Paul S. 2017. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, **38**(4), 13–26.

[37] Larue, Othalia, West, Robert, Rosenbloom, Paul S, Dancy, Christopher L, Samsonovich, Alexei V, Petters, Dean, and Juvina, Ion. 2018. Emotion in the common model of cognition. *Procedia computer science*, **145**, 740–746.

[38] Lebiere, Christian. 1999. The dynamics of cognitive arithmetic. *Kognitionswissenschaft Special issue on cognitive modelling and cognitive architectures, D. Wallach and H. A. Simon (eds.)*, **8**, 5–19.

[39] Lebiere, Christian, Cranford, Edward, Aggarwal, Palvi, Cooney, Sarah, Tambe, Milind, and Gonzalez, Cleotilde. 2023. Cognitive Modeling for Personalized, Adaptive Signaling for Cyber Deception. Pages 59–82 of: *Bao, T., Tambe, M., Wang, C. (eds) Cyber Deception. Advances in Information Security, vol 89*. Springer.

[40] MacLeod, Colin M. 1991. Half a century of research on the Stroop effect: an integrative review. *Psychological bulletin*, **109**(2), 163.

[41] Marcílio, Wilson E, and Eler, Danilo M. 2020. From explanations to feature selection: assessing SHAP values as feature selection mechanism. Pages 340–347 of: *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*. Ieee.

[42] Mitsopoulos, Konstantinos, Somers, Sterling, Schooler, Joel, Lebiere, Christian, Pirolli, Peter, and Thomson, Robert. 2021. Toward a psychology of deep reinforcement learning agents using a cognitive architecture. *Topics in Cognitive Science*, **14**(4), 756–779.

[43] Moreno, Lorena, Briñol, Pablo, and Petty, Richard E. 2022. Metacognitive confidence can increase but also decrease performance in academic settings. *Metacognition and Learning*, **17**(1), 139–165.

[44] Nelson, Thomas O, and Narens, L. 1990. Metamemory: A theoretical framework and new findings. Pages 125–173 of: *Psychology of learning and motivation*, vol. 26. Elsevier.

[45] Newell, Allen. 1973. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. Page 283–308 of: *Visual information processing*. Elsevier.

[46] Newell, Allen. 1990. *Unified Theories of Cognition*. Harvard University Press, USA.

[47] O'Reilly, Randall C, and Munakata, Yuko. 2000. *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT press.

[48] Parrish, A.E., and Brosnan, S.F. 2012. Primate cognition. Pages 174–180 of: *Encyclopedia of Human Behavior (Second Edition)*. Elsevier.

[49] Paynter, Christopher A, Reder, Lynne M, and Kieffaber, Paul D. 2009. Knowing we know before we know: ERP correlates of initial feeling-of-knowing. *Neuropsychologia*, **47**(3), 796–803.

[50] Petty, Richard E, Briñol, Pablo, Tormala, Zakary L, and Wegener, Duane T. 2007. The role of metacognition in social judgment. *Social psychology: Handbook of basic principles*, **2**, 254–284.

[51] Popper, Karl R. 1968. Epistemology without a knowing subject. Pages 333–373 of: *Studies in Logic and the Foundations of Mathematics*, vol. 52. Elsevier.

[52] Pouget, Alexandre, Drugowitsch, Jan, and Kepecs, Adam. 2016. Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, **19**(3), 366–374.

[53] Ptasczynski, Lena Esther, Steinecker, Isa, Sterzer, Philipp, and Guggenmos, Matthias. 2022. The value of confidence: Confidence prediction errors drive value-based learning in the absence of external feedback. *PLOS Computational Biology*, **18**(10), e1010580.

[54] Rao, Rajesh PN, and Ballard, Dana H. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, **2**(1), 79–87.

[55] Rizzolatti, Giacomo, and Craighero, Laila. 2004. The mirror-neuron system. *Annu. Rev. Neurosci.*, **27**, 169–192.

[56] Rosenbloom, Paul S., Laird, John E., Lebiere, Christian, Stocco, Andrea, Granger, Richard H., and Huyck, Christian. 2024. A Proposal for Extending the Common Model of Cognition to Emotion. In: *Proceedings of the 22nd Annual Meeting of the International Conference on Cognitive Modeling*.

[57] Rummel, Jan, and Meiser, Thorsten. 2013. The role of metacognition in prospective memory: Anticipated task demands influence attention allocation strategies. *Consciousness and Cognition*, **22**(3), 931–943.

[58] Scheck, Petra, and Nelson, Thomas O. 2005. Lack of pervasiveness of the underconfidence-with-practice effect: boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General*, **134**(1), 124.

[59] Schneider, Wolfgang, and Lockl, Kathrin. 2008. Procedural metacognition in children: Evidence for developmental trends. *Handbook of metamemory and memory*, **14**, 391–409.

[60] Schoenherr, Jordan R, Leth-Steensen, Craig, and Petrusic, William M. 2010. Selective attention and subjective confidence calibration. *Attention, Perception, & Psychophysics*, **72**(2), 353–368.

[61] Schoenherr, Jordan Richard, and Lacroix, Guy L. 2020. Performance monitoring during categorization with and without prior knowledge: A comparison of confidence calibration indices with the certainty criterion. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, **74**(4), 302–315.

[62] Schoenherr, Jordan Richard, and Thomson, Robert. 2021. Persuasive features of scientific explanations: Explanatory schemata of physical and psychosocial phenomena. *Frontiers in Psychology*, **12**, 644809.

[63] Schultz, Wolfram, Dayan, Peter, and Montague, P Read. 1997. A neural substrate of prediction and reward. *Science*, **275**(5306), 1593–1599.

[64] Schunn, Christian D, Reder, Lynne M, Nhouyvanisvong, Adisack, Richards, Daniel R, and Stroffolino, Philip J. 1997. To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **23**(1), 3.

[65] Shea, Nicholas, Boldt, Annika, Bang, Dan, Yeung, Nick, Heyes, Cecilia, and Frith, Chris D. 2014. Supra-personal cognitive control and metacognition. *Trends in cognitive sciences*, **18**(4), 186–193.

[66] Shekhar, Medha, and Rahnev, Dobromir. 2021. Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, **25**(1), 12–23.

[67] Spada, Marcantonio M, Nikčević, Ana V, Moneta, Giovanni B, and Wells, Adrian. 2008. Metacognition, perceived stress, and negative emotion. *Personality and Individual Differences*, **44**(5), 1172–1181.

[68] Sun, Ron. 2016. *Anatomy of the mind: exploring psychological mechanisms and processes with the Clarion cognitive architecture*. Oxford University Press.

[69] Taatgen, Niels A. 2017. Cognitive architectures: Innate or learned? In: *AAAI Technical Report FS-17-05*.

[70] Thomson, Robert, and Frangia, William. 2023. Investigating the Use of Belief-Bias to Measure Acceptance of False Information. Pages 149–158 of: *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer.

[71] Thomson, Robert, Cranford, Edward, Somers, Sterling, and Lebiere, Christian. 2024. A Novel Approach to Intrusion Detection Using a Cognitively-Inspired Algorithm. In: *Proceedings of the 57th Hawaii International Conference on System Sciences*.

[72] Vinokurov, Yury, Lebiere, Christian, Herd, Seth, and O'Reilly, Randall. 2011. A metacognitive classifier using a hybrid ACT-R/Leabra architecture. In: *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[73] Vinokurov, Yury, Lebiere, Christian, Wyatte, Dean, Herd, Seth, and O'Reilly, Randall. 2012. Unsurpervised learning in hybrid cognitive architectures. In: *Workshops at the twenty-sixth AAAI conference on artificial intelligence*.

[74] Yeung, Nick, and Summerfield, Christopher. 2012. Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**(1594), 1310–1321.

[75] Yeung, Nick, Botvinick, Matthew M, and Cohen, Jonathan D. 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review*, **111**(4), 931.