Towards a Computational Cognitive Neuroscience Model of Creativity

Hugo Chateau-Laurent, Frédéric Alexandre

Towards a Computational Cognitive Neuroscience Model of Creativity

Hugo Chateau-Laurent
Mnemosyne, Inria Bordeaux Sud-Ouest
LaBRI, CNRS, Université de Bordeaux
Institut des Maladies Neurodégénératives
CNRS, Université de Bordeaux, France
hugo.chateau-laurent@inria.fr

Frédéric Alexandre
Mnemosyne, Inria Bordeaux Sud-Ouest
LaBRI, CNRS, Université de Bordeaux
Institut des Maladies Neurodégénératives
CNRS, Université de Bordeaux, France
frederic.alexandre@inria.fr

***Abstract*—Recent progress in AI has expanded the boundaries of the cognitive functions that can be simulated, but creativity remains a challenge. Neuroscience sheds light on its mechanisms and its tight relationship with episodic memory and cognitive control, while machine learning provides preliminary models of these mechanisms. We present these lines of research and explain how they can be exploited in the domain of computational creativity in order to further expand the capabilities of AI.**

***Index Terms*—creativity, explicit memory, episodic memory, prospective memory, cognitive control, imagination, computational cognitive neuroscience, artificial intelligence**

I. Introduction

Creativity is a fundamental feature of human information processing but a historically challenging task for a machine. It is generally agreed that it includes two defining characteristics, namely coming up with a novel idea and verifying that this idea is appropriate for the task at play [1]. They are respectively termed divergent and convergent thinking [2].

Despite tremendous progress made with machine learning, artificial intelligence models suffer from a series of weaknesses including but not limited to overfitting and poor transferability [3]. Thinking through analogy and imagination would be key ingredients for solving these problems. Unfortunately, such creative functions seem to remain a challenge for the machine. Yet, it is declared in [4] that "creativity isn't magical. It's an aspect of normal human intelligence". In accordance with M. Boden, we argue that creativity results from the functioning of the human brain and should consequently be better understood by studying its physical substrate. As a consequence, we propose to relate computational creativity (CC) to computational cognitive neuroscience (CCN). It has already been suggested in [5] that current modeling work in CCN carried out to remedy current limitations of artificial intelligence could be directly beneficial to CC by unravelling the brain mechanisms responsible for divergent and convergent thinking. Here we propose to extend this view by providing more details on recent experiments in neuroscience and recent modeling efforts in that direction. In that perspective, we briefly sketch the mnemonic architecture of the human brain, going in more details on mechanisms reported to be involved in imagination and cognitive control, participating in creativity. Then we report recent modeling efforts to implement these mechanisms and discuss how they might be adapted to CC.

II. The human mnemonic architecture

A. Implicit memory

In cognitive neuroscience, a major distinction is made between implicit and explicit memory, also called non-declarative and declarative memory respectively [6]. The former corresponds to the slow learning of regularities present in the world. This learning is well captured in current machine learning models, as they rely on statistical methods. This is for example the case with the layered architecture of deep networks [7] that has demonstrated high performance in pattern identification and is sometimes compared to the architecture of the sensory cortex. It is also the case with the most classical approaches of reinforcement learning, the actor-critic architectures [8], which are today among the best models for action selection and decision making. These approaches are paralleled with the loops between the basal ganglia and the most ancient parts of the frontal cortex, namely the motor and premotor cortex (standing for the actor) and the orbitofrontal cortex (standing for the critic). These models perform implicit learning because they simply learn from a statistical analysis of trials and errors, regularities in labelled patterns, rewarding value of states or sensorimotor associations, with no explicit knowledge of the world structure [3]. After learning, these values cached in weights are exploited implicitly to respond to new cases. As it can be seen with the recent success of machine learning, these approaches are very successful when large training corpora are available, but they suffer from several weaknesses, precisely because their learning is implicit. Two sets of weaknesses have been studied in the context of two different mechanisms that might remedy them, both corresponding to building an explicit memory: episodic learning and cognitive control.

B. Explicit memory: episodic learning

A first set of weaknesses is observed in deep networks, where learning takes place in a layered architecture with distributed, integrated and overlapping representations. This is the reason why these models can generalize, at the price of a

very slow learning. Conversely, they are not good for learning data without structure nor regularities. They learn slowly, from global aspects to details, and they forget the individual cases they have received. They are also sensitive to catastrophic forgetting: if they first learn to identify a certain pattern and then learn to identify another one, they will forget the first kind of identification if they don't have a regular recall of the previously learnt association, thus adding to the slowness of learning.

It is explained in [9] how the hippocampus which implements episodic learning could complement the cortex in order for the mammalian brain to remedy that kind of weaknesses. This cerebral structure in the medial temporal lobe binds information from most regions of the cortex and is able to learn the current cortical activation in one shot. Due to mechanisms of sustained activity in the cortex evoked below, this activation stands for the current episode processed by the cortex, hence the term episodic learning. This memory is explicit because, in contrast to the cortex that evokes relations acquired from many episodes, this learning manipulates individual episodes explicitly. This is associated with specific characteristics of hippocampal information processing: an elaborated function of pattern separation based on sparse coding [10] to minimize interference when learning a new episode and a mechanism of recall enabling the hippocampus to rebuild an originally stored pattern from a partial cue thanks to its unique recurrent architecture. Recall is associated with a mechanism called replay [11] to reactivate a stored episode in the cortex. This process helps solving the problem of catastrophic interference because, when you are learning a new ability in the cortex, it has been shown that during sleep, the hippocampus replays other kinds of examples to balance learning [11]; and what is replayed here are explicit and unique episodes and not a statistical implicit model. This illustrates an interesting duality in our memories between knowledge (I know that) and skill (I know how).

C. Explicit memory: cognitive control

Let us now introduce another set of weaknesses associated with implicit learning. Reinforcement learning models are very powerful in stable worlds and can assign a value to states and actions and guide behavior accordingly, with no explicit knowledge of the rules of the world. This is no longer the case if the behavior to be learnt is context-dependent, or in other words when the rules change depending on certain perceptual cues or on internal information (instructions given or simply emotions or motivations). An implicit model could unlearn the old behavior and learn the new one but this would come at a cost of time. This is not compatible with the flexibility of human behaviour that can exhibit rapid changes from one situation to an other. In addition, we can see that this kind of change is explicit and deliberative. In this case, the solution would be to not only rely on the default behavior proposed by the implicit model from external information but to explicitly identify the context, inhibit the default behavior and instead promote the appropriate one which has been stored from previous learning. This is what we call cognitive control: whereas simple decision making relies on perception to control actions and promote a certain behavior, cognitive control relies on external and internal contextual cues to control behavior in a top-down way.

It has long been known in neuroscience that the prefrontal cortex (PFC) is the structure responsible for this function [12]. This structure, particularly developed in humans (30% of the total surface of the cortex), is characterized by its ability to show a long-lasting sustained activity, also called working memory. In certain circumstances, it is able to inhibit a maladapted default behavior and promote a more appropriate one for an extended period of time. This is done by biasing the activity of the cortex, i.e. by artificially imposing an increased activity to some sensory regions of the cortex through an attentional process [13]. In this way, a non-dominant behavior can become eligible and active. When rule changes, the working memory can be updated to promote another behavior. Deciding if the rules have changed (from the monitoring of errors measured between the anticipated and observed outcomes) and learning what contextual cues to maintain in working memory are the roles of the medial and lateral PFC respectively [14]. In these regions, the neuronal populations are organized according to the nature of the task [15]. It has been shown that they are organized along several dimensions on the cortical surface [16]. From rostral to more caudal portions, more and more abstract rules can be observed. There is also a dorsal-ventral axis, with rules predominantly concerned with the semantic meaning of the task on the ventral side and rules controlling the syntactic organization of the task on the dorsal side. When errors indicate that the rule has changed, it is possible to immediately test other rules which have been proven efficient for this task, else to look for rules applied in similar (i.e. sharing characteristics) tasks. Another proposed strategy is to build new rules by mixing old ones [17]. As discussed below, these are key mechanisms for creativity.

III. Deliberating in the future

A. Prospective memory

We have explained that the hippocampus is able to bind distributed activities of the cortex to explicitly create a memory of a specific episode and that the PFC is able to bias the activity of the cortex to inhibit the default implicit behavior and explicitly promote a behavior more appropriate in the current context [18]. What is fundamental to understand here is that the hippocampus also receives pieces of information from the PFC and binds them within episodes which are also sent back to the PFC thus participating to its training and activation. Reciprocally, the PFC receives contextual information elaborated in the hippocampus and can bias the activity of the hippocampus in order to control both the storage and retrieval of episodes [19]. Furthermore, it is now widely accepted that memory is a (re)constructive process, and that nearly the same circuitry is used for not only remembering (or reconstructing) the past, but also imagining the future [20]. In summary,

the hippocampus and the PFC both form a generative model by learning an explicit model of the world. By sampling from various episodes and anticipating the consequences of one's actions, they are able to internally simulate and evaluate candidate strategies. This is what we call imagination and what is realized in the process of thinking [21].

B. What about creativity ?

Neuroscientific studies about creativity are in accordance with the general picture given here and provide additional details. [22] report an interesting fMRI experiment demonstrating the role of the medial PFC and hippocampus in reorganizing memory when new information is presented. In the experiment, a series of narrative videos is shown, including clues linking some of them and stimulating insight in participants (i.e. the sudden discovery of a solution to a problem). It is observed that these moments are associated with the emergence of new mnemonic representations while irrelevant events are pruned out. The paper reports individual events activating the hippocampus and mismatch response in the anterior hippocampus being sent to the medial PFC. [23] report the role of the medial PFC in inhibiting automatic common ideas through its dorsal part and allowing for flexibility and original ideas generation through its ventral part. The medial PFC can detect the co-existence of several solutions including non-dominant solutions and can switch attention to favor one of them [24], with the level of originality in creativity rather linked to broadened and not focused attention [25]. Concerning divergent thinking, insight is defined as the reinterpretation of a situation to produce non-dominant interpretation [24] and characterized by the activation of the medial PFC (for conflict monitoring) and of the medial temporal lobe including the hippocampus (associated with semantic integration before insight). In that paper, the role of mood is also underlined, associated with the broadening of semantic processing. This might correspond to the difference between spontaneous creativity monitored by such bottom-up information, opposed to deliberate creativity monitored in a top-down way by the PFC [1]. Concerning convergent thinking, verifying the semantic and syntactic pertinence of a newly proposed behavior is part of the top-down control of the PFC observed during creativity, shown to activate the ventral and dorsal regions respectively [2]. Interestingly, in prospective memory, this verification can be done internally using the learnt generative models, with no need for real action.

IV. Computational aspects

A. Models of explicit memory

Whereas recent achievements in machine learning are mainly related to models of implicit memory, more attention has recently been given to models of explicit memory in order to provide more realistic and flexible models. In reinforcement learning, this is for example the case of episodic and meta reinforcement learning [26], introduced to accelerate deep reinforcement learning with the same kind of arguments as provided above. In the case of supervised learning, the principle of conceptors [27] has also been introduced as an attentional process by restricting processing to a sub-region of the data space. Beyond these computational principles, other models have a deeper biological grounding. Let us first mention a model of the hippocampus able to store and recall episodes [28], with the emphasis set on interactions between various inner mechanisms. Particularly, in addition to the learning of specific episodes, this model is also able to discover statistical regularities in the episodes, which is shown to be a potential substrate for transitive inference (i.e. the discovery that if A is associated to B and B is associated to C, then A is associated to C), an important ingredient of creativity. Other computational studies have extended the traditional reinforcement learning paradigm to account for hippocampal sequence generation. For example, a recently proposed model provides an explanation as to why and how the hippocampus is capable of not only replaying experienced sequences of states, but "imagining" unexperienced ones as well [29]. Concerning cognitive control, [14] implements a neuronal model based on the principles of predictive coding. Errors detected between the anticipated and actual outcomes are mapped and associated with their context to subsequently trigger a more adapted behavior when the same context is detected.

B. What about computational creativity ?

To our knowledge, there are presently very few CCN models dealing explicitly with creativity, with the notable exceptions of [17] and [30]. The former model mainly focuses on the action of the medial PFC in creativity. It proposes a biologically informed computational account of how we can flexibly switch from one set of rules to another and explains that we might create a new set of rules from combining existing ones. In this model using simple symbolic tasks, the role of the hippocampus is not considered, which leaves considerable room for improvement. In the latter study, a computational account of creativity emphasizing the importance of binding information from multiple sources is given. In our opinion, what is currently lacking is a computational model of the interactions between the hippocampus (arbitrary binding and episodic learning) and the PFC (control of encoding, recall and imagination) dedicated to creativity. Beyond implementing controlled episodic memory, we believe that such a model would shed light on the ability of the "construction system of the brain" [31] to creatively think about novel situations and verifying their task relevance (Fig. 1).

V. Concluding remarks

In this paper, we attempted to lay down the foundations for a CCN approach to creativity based on the combination of cognitive control and episodic learning. We have explained that several models already propose mechanisms participating to creative processes in the brain but several characteristics need to be studied in more depth. For example, the role of hemispheric differences is often mentioned in neuroscientific experiments about creativity [24] whereas this concept of

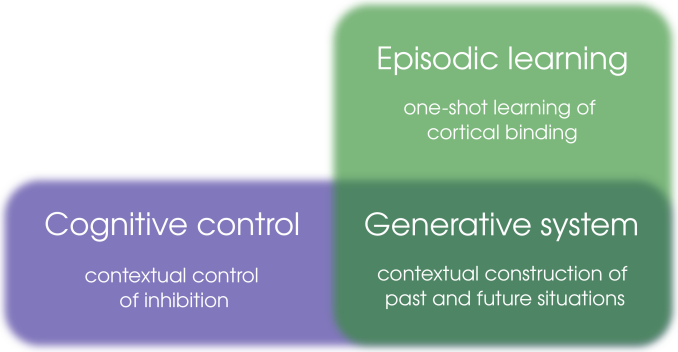

Fig. 1. The proposed framework combines cognitive control and episodic learning for solving flexible episodic memory and creative planning.

lateralization is mostly absent in models. We believe that an accurate model of how creativity is implemented in the brain would explain experimental findings, such as the deactivation of the lateral PFC and activation of the medial PFC that have been observed during music improvisation [2]. We can hypothesize that the medial PFC is responsible for tracking errors in the improvisation, while the constrained control of behavior implemented by the lateral PFC is being inhibited. We also believe that better understanding the computational mechanisms associated to creativity would greatly augment machine intelligence considering their role in imagination, planning, decision-making, navigation, and arguably all other cognitive abilities. The development of a detailed computational framework will likely lead to the demystification of creativity as a unique human capability. Even more striking is the possibility that creativity might be reduced to certain aspects of control, memory and decision making.

REFERENCES

[1] A. Dietrich, "The cognitive neuroscience of creativity," *Psychonomic Bulletin & Review*, vol. 11, no. 6, pp. 1011–1026, Dec. 2004. [Online]. Available: https://doi.org/10.3758/BF03196731

[2] R. E. Jung, B. S. Mead, J. Carrasco, and R. A. Flores, "The structure of creative cognition in the human brain," *Frontiers in Human Neuroscience*, vol. 7, 2013. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2013.00330/full

[3] G. Marcus, "Deep learning: A critical appraisal," *arXiv preprint arXiv:1801.00631*, 2018.

[4] M. A. Boden, "Computer Models of Creativity," *AI Magazine*, vol. 30, no. 3, pp. 23–23, Jul. 2009, number: 3. [Online]. Available: https://www.aaai.org/ojs/index.php/aimagazine/article/view/2254

[5] F. Alexandre, "Creativity explained by computational cognitive neuroscience," in *Proceedings of the Eleventh International Conference on Computational Creativity, ICCC 2020, Coimbra, Portugal, September 7-11, 2020*, F. A. Cardoso, P. Machado, T. Veale, and J. M. Cunha, Eds. Association for Computational Creativity (ACC), 2020, pp. 374–377. [Online]. Available: http://computationalcreativity.net/iccc20/papers/119-iccc20.pdf

[6] L. R. Squire, "Declarative and nondeclarative memory: multiple brain systems supporting learning and memory." *Journal of cognitive neuroscience*, vol. 4, no. 3, pp. 232–243, 1992. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/23964880

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: http://dx.doi.org/10.1038/nature14539

[8] D. Joel, Y. Niv, and E. Ruppin, "Actor–critic models of the basal ganglia: new anatomical and computational perspectives," *Neural Networks*, vol. 15, no. 4-6, pp. 535–547, Jul. 2002. [Online]. Available: http://dx.doi.org/10.1016/s0893-6080(02)00047-3

[9] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory." *Psychological review*, vol. 102, no. 3, pp. 419–457, Jul. 1995. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/7624455

[10] R. Kassab and F. Alexandre, "Pattern separation in the hippocampus: distinct circuits under different conditions," *Brain Structure & Function*, vol. 223, no. 6, pp. 2785–2808, 2018.

[11] I. Stoianov, D. Maisto, and G. Pezzulo, "The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning," *bioRxiv*, p. 2020.01.16.908889, Jan. 2020. [Online]. Available: https://www.biorxiv.org/content/10.1101/2020.01.16.908889v1

[12] J. Fuster, *The prefontal cortex. Anatomy, physiology and neurophysiology of the frontal lobe*. Raven Press, New-York, 1989.

[13] R. C. O'Reilly, D. C. Noelle, T. S. Braver, and J. D. Cohen, "Prefrontal cortex and dynamic categorization tasks: representational organization and neuromodulatory control." *Cereb Cortex*, vol. 12, no. 3, pp. 246–257, Mar. 2002. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11839599

[14] W. H. Alexander and J. W. Brown, "Hierarchical error representation: a computational model of anterior cingulate and dorsolateral prefrontal cortex," *Neural Computation*, vol. 27, no. 11, pp. 2354–2410, 2015.

[15] P. Domenech and E. Koechlin, "Executive control and decision-making in the prefrontal cortex," *Current Opinion in Behavioral Sciences*, vol. 1, pp. 101–106, Feb. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.cobeha.2014.10.007

[16] R. C. O'Reilly, "The What and How of prefrontal cortical organization," *Trends in Neurosciences*, vol. 33, no. 8, pp. 355–361, Aug. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.tins.2010.05.002

[17] A. Collins and E. Koechlin, "Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making," *PLOS Biology*, vol. 10, no. 3, pp. e1 001 293+, Mar. 2012. [Online]. Available: http://dx.doi.org/10.1371/journal.pbio.1001293

[18] J. D. Cohen and R. C. O'Reilly, "A Preliminary Theory of the Interactions Between Prefrontal Cortex and Hippocampus that Contribute to Planning and Prospective Memory," in *Prospective Memory: Theory and Applications*, M. Brandimonte, G. O. Einstein, and M. A. McDaniel, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, 1996, pp. 267–296.

[19] H. Eichenbaum, "Prefrontal–hippocampal interactions in episodic memory," *Nature Reviews Neuroscience*, vol. 18, no. 9, pp. 547–558, 2017.

[20] D. L. Schacter, D. R. Addis, and R. L. Buckner, "Remembering the past to imagine the future: the prospective brain," *Nature Reviews Neuroscience*, vol. 8, no. 9, pp. 657–661, Sep. 2007. [Online]. Available: https://www.nature.com/articles/nrn2213

[21] G. Pezzulo and C. Castelfranchi, "Thinking as the control of imagination: a conceptual framework for goal-directed systems," *Psychological Research PRPF*, vol. 73, no. 4, pp. 559–577, 2009. [Online]. Available: http://dx.doi.org/10.1007/s00426-009-0237-z

[22] B. Milivojevic, A. Vicente-Grabovetsky, and C. Doeller, "Insight Reconfigures Hippocampal-Prefrontal Memories," *Current Biology*, vol. 25, no. 7, pp. 821–830, Mar. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982215000652

[23] N. Mayseless, A. Eran, and S. G. Shamay-Tsoory, "Generating original ideas: The neural underpinning of originality," *NeuroImage*, vol. 116, pp. 232–239, Aug. 2015.

[24] J. Kounios and M. Beeman, "The cognitive neuroscience of insight," *Annual Review of Psychology*, vol. 65, pp. 71–93, 2014.

[25] H. Takeuchi, Y. Taki, R. Nouchi, R. Yokoyama, Y. Kotozaki, S. Nakagawa, A. Sekiguchi, K. Iizuka, S. Hanawa, T. Araki, C. M. Miyauchi, K. Sakaki, Y. Sassa, T. Nozawa, S. Ikeda, S. Yokota, D. Magistro, and R. Kawashima, "Originality of divergent thinking is associated with working memory–related brain activity: Evidence from a large sample study," *NeuroImage*, vol. 216, p. 116825, Aug. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1053811920303128

[26] M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, and D. Hassabis, "Reinforcement Learning, Fast and Slow," *Trends in Cognitive Sciences*, vol. 23,

no. 5, pp. 408–422, May 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364661319300610

[27] X. He and H. Jaeger, "Overcoming Catastrophic Interference using Conceptor-Aided Backpropagation," *ICLR*, Feb. 2018. [Online]. Available: https://openreview.net/forum?id=B1al7jg0b

[28] A. C. Schapiro, N. B. Turk-Browne, M. M. Botvinick, and K. A. Norman, "Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 372, no. 1711, Jan. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124075/

[29] M. Khamassi and B. Girard, "Modeling awake hippocampal reactivations with model-based bidirectional search," *Biological Cybernetics*, vol. 114, no. 2, pp. 231–248, Apr. 2020. [Online]. Available: https://doi.org/10.1007/s00422-020-00817-x

[30] P. Thagard and T. C. Stewart, "The aha! experience: Creativity through emergent binding in neural networks," *Cognitive science*, vol. 35, no. 1, pp. 1–33, 2011.

[31] D. Hassabis and E. A. Maguire, "The construction system of the brain," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1521, pp. 1263–1271, 2009.