DeepThought: An Architecture for Autonomous Self-motivated Systems

Arlindo L. Oliveira
INESC-ID & ELLIS Lisbon Unit
Instituto Superior Técnico, ULisboa, Portugal
arlindo.oliveira@tecnico.ulisboa.pt

Tiago Domingos
MARETEC
Instituto Superior Técnico, ULisboa, Portugal
tdomingos@tecnico.ulisboa.pt

Mário Figueiredo
Inst. Telecomunicações & ELLIS Lisbon Unit
Instituto Superior Técnico, ULisboa, Portugal
mario.figueiredo@tecnico.ulisboa.pt

Pedro U. Lima
ISR & ELLIS Lisbon Unit
Instituto Superior Técnico, ULisboa, Portugal
pedro.lima@tecnico.ulisboa.pt

Abstract

The ability of large language models (LLMs) to engage in credible dialogues with humans, taking into account the training data and the context of the conversation, has raised discussions about their ability to exhibit intrinsic motivations, agency, or even some degree of consciousness. We argue that the internal architecture of LLMs and their finite and volatile state cannot support any of these properties. By combining insights from *complementary learning systems*, *global neuronal workspace*, and *attention schema* theories, we propose to integrate LLMs and other deep learning systems into an architecture for cognitive language agents able to exhibit properties akin to agency, self-motivation, even some features of meta-cognition.

1 Introduction and motivation

The success of LLMs, based on the transformer architecture [1] and trained on very large datasets [2, 3], has raised significant questions about whether such systems could exhibit some characteristics of higher-level natural intelligence: agency, self-motivation, even some aspects of consciousness. In fact, when equipped with long contexts, LLMs can engage in dialogues where their utterances can easily be mistaken for true signs of agency or sentience. For the purposes of this paper, we consider an LLM to be an autoregressive language model that predicts the next token by computing its conditional probability given the tokens in the context. More formally, language models learn a distribution $P(w_i \mid w_{i-1}, ..., w_{i-c})$, where c is the context length. The model can then be used by sampling from the learned distribution, one token at a time. The use of *reinforcement learning from human feedback* (RLHF) or fine-tuning techniques does not change the fundamental properties of language models, although it can significantly improve their performance. Given the sometimes surprising emergent properties of LLMs [4], there has been significant disagreement on whether scaling up these models may lead to the emergence of agency and purpose. In this position paper, we argue that simply scaling up LLMs is unlikely to lead to the emergence of the hallmarks properties of true intelligence, namely self-motivation and agency. However, integrating LLMs with other neural network-based architectures for vision and interaction with the external world as modules of a larger system may yield much richer behaviors that, if observed in humans or animals, are described as attention, agency, self-motivation, and even consciousness.

Agency refers to the capacity of an entity to act intentionally, make choices, and exert control over its actions and decisions [5]. An agent has the power to make meaningful choices that influence its state

and that of the world around it. Agency is therefore necessarily associated with entities with internal states (memory) and capable of making intentional, self-motivated choices based on those states.

LLMs have only a limited state, corresponding to the sequence of tokens in the context. The internal weights and other parameters do not store state information, as they are fixed after training and during interaction with the model. Since the context is finite, it provides only a limited state. Furthermore, every time an LLM starts a new conversation, it has an empty context, thereby ignoring all previous conversations. It is thus clear that LLMs, even if scaled up to include much more training data, cannot, by themselves, exhibit the features that characterize agency and intrinsically motivated systems.

Agency implies the ability of a system (the agent) to act purposefully and make choices to achieve goals. It thus implies a degree of autonomy and the ability to decide based on some form of cognition and processing. While the criteria for agency can vary depending on the context and perspective, there are several characteristics often associated with agents, including: autonomy (1), sensing (2), information processing (3), goal-oriented behavior (4), decision-making (5), adaptability (6), ability to act (7), persistence (8), learning (9), and intent or purpose (10). LLMs, by themselves, being essentially stateless systems driven by prompts, clearly lack several of these properties, although they possess some to a degree. Most notably, they lack properties 1, 4, 8, 9, and 10, even though their ability to dialogue may simulate, in some cases, some of them. The fact that their weights are fixed after training means they are not able to learn from interactions, thereby lacking property 9 (although they exhibit some form of in-context learning, it vanishes when each conversation ends). However, they are able to perform some limited sensing of the environment (by reading the input tokens), adapt to the input received so far (by considering the tokens in the context), make decisions based on the inputs, and perform a limited set of actions (by generating output tokens). To endow systems with the missing abilities from this list, it is necessary to enrich these models with memory, the ability to adapt to new situations, and mechanisms to pursue goals.

2 Language and Cognitive Language Agents

The ever-increasing popularity and capabilities of LLMs [2, 3] led rapidly to the idea that they could be the central components of architectures for autotelic agents. Well-known cognitive architectures, such as SOAR [6], which have been proposed and tested for several decades, can gain significant flexibility and power by incorporating language models as components. In fact, the use of language models addresses two key limitations of cognitive architectures: they are limited to domains that can be formally specified; they require many hand-coded rules [7]. LLMs are a natural choice to address these two limitations since they operate in arbitrary settings and are extremely flexible.

In the most common usage to date, language models take text as input and generate text as output. The only state they maintain corresponds to the context, the limited history of the interaction with the user, possibly augmented by an initial prompt that guides the behavior. A more sophisticated use of LLMs places them in a feedback loop with the outside world by converting observations into text and using the LLMs' output to choose actions, thus creating language agents [7]. More advanced architectures, known as cognitive language agents, add additional components such as goal prioritization mechanisms (e.g., AutoGPT [8]), code generation and storage [9], memory, and other symbolic reasoning abilities [10]. For instance, generative agents [11] use an architecture that includes a memory stream and a retrieval mechanism that, when coupled with the abilities of GPT-4, leads to agency-revealing behavior that resembles that of human beings.

Several of these proposals have shown that cognitive language agents exhibit relevant properties that are characteristic of agency, including autonomy, sensing, goal-oriented and autotelic behavior, and adaptability. However, we believe they fall short of achieving the full potential of self-motivated agents because the goals they pursue are either imposed externally or the result of random exploration of the state space and they lack attention mechanisms and meta-cognition (the ability to introspect their own internal cognition mechanisms).

We argue that meta-cognition and autotelic behavior require the use of *attention*, *i.e.*, the ability to focus the available resources on specific parts of the external input or of the internal state. This possibility to focus on (attend to) specific sub-parts of the available data is also behind our ability to break long-term goals into self-generated short-term goals and to adopt the behavior that leads to the fulfillment of the resulting sub-goals. Our proposal is inspired by a number of theories of attention and conscious behavior.

3 Theories of attention and conscious behavior

Complementary learning systems theory (CLS) [12, 13] is a cognitive neuroscience theory that proposes a framework for understanding how the human brain processes and stores information, particularly in the context of memory and learning. This theory suggests that the brain employs two distinct and complementary systems for memory and learning: the hippocampus and the neocortex.

Rapid learning takes place in the hippocampus, in particular in the *medium temporal lobe* (MTL), which is thought to be responsible for fast, episodic, and one-shot learning. It quickly encodes new information and experiences and plays a key role in declarative memory. This system also helps in differentiating and encoding similar but distinct experiences, preventing interference between them. The hippocampus adapts rapidly to new learning situations and is crucial for the formation of new episodic memories. Slow learning happens in the neocortex, which is associated with slower, more gradual learning. This system consolidates information over time and is responsible for the storage and retrieval of semantic or factual knowledge, and well-established memories. It is also involved in extracting commonalities and general principles from individual experiences.

CLS theory argues that these two systems work together in a complementary manner to support memory and learning. Initially, when we encounter new information, the hippocampus rapidly encodes it, but these memories are fragile and degrade over time. The neocortex gradually consolidates these memories, making them more stable and resistant to interference. A key idea behind the CLS theory is that the rapid learning facilitated by the hippocampus is essential for adapting to novel situations, while the slower, more stable neocortical learning is responsible for building a reliable knowledge base. This duality allows humans to acquire new knowledge quickly and integrate it into their existing understanding of the world, something sorely lacking in current deep learning systems. Recently, McClelland et al. [14] suggested an architecture for the *brain understanding system*, depicted in Fig. 1 (left). While this figure aims to represent the fundamental architecture of biological brains, it can also serve as a blueprint for autotelic artificially intelligent systems.
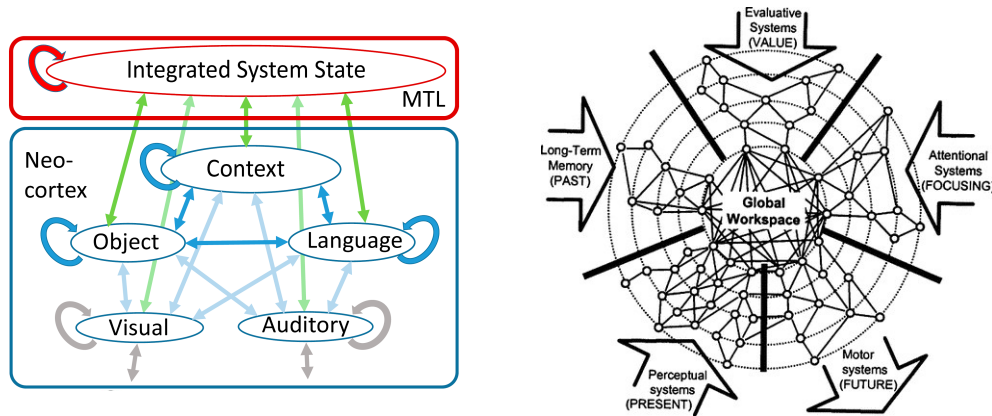

Figure 1: Left: The brain understanding system (reproduced from [14]). The red arrow represents fast-learning connections while the green arrows represent slow-learning connections. Right: an illustration of the global workspace theory (reproduced from [15]). When information is brought into the global workspace, it becomes conscious and is broadcast to all the (unconscious) processors.

Global workspace theory (GWT) (which is closely related to the *global neuronal workspace* – GNW – theory) [15, 16, 17] addresses distinct aspects of cognitive processing, by attempting to explain how the brain processes information and achieves conscious awareness. While CLS theory primarily deals with the organization of memory systems and how different brain regions work together to support memory and learning, GWT/GNW is a theory of consciousness and attention, primarily concerned with how information is selected and broadcast to conscious awareness. GWT/GNW postulates a "global workspace" in the brain where information from various cognitive processes (perception, memory retrieval, decision-making, etc.) competes for attention and conscious processing. In GWT/GNW, information from different sources is integrated and made available to conscious awareness (Figure 1, right). GNW postulates that, apart from specialized cortical areas that process different types of information, there is a distributed network of GNW neurons with long-range

connections, which selectively influence specific processing neurons and act as a communication hub, receiving input from lower levels and transmitting guidance to different processing regions.

Although they have different assumptions and postulate different mechanisms, CLS and GWT/GNW are clearly related. This relationship suggests that combining their ideas can provide a more comprehensive account of cognitive processing, especially in the context of memory and conscious awareness. In particular, GWT/GNW may help explain how the information processed and consolidated by the CLS's neocortical system becomes accessible to conscious awareness. When a memory or piece of knowledge is retrieved and brought into the global workspace, it can be consciously experienced. GWT/GNW's notion of attention and global workspace can also be related to CLS in terms of how attention processes affect learning and memory encoding. Information that receives attention is more likely to be rapidly encoded by the hippocampus, initiating or strengthening the learning process.

The role of the MTL in CLS theory could, therefore, be somehow related to the role of the global workspace in GWT/GNW. The fact that the mechanisms proposed by the two theories do not fully coincide may be due to our limited knowledge of the process but does not preclude the existence of a deep relationship between the two postulated mechanisms. In fact, the MTL may well play the role of a sensory hub where visual features are "bound" into single, conscious (reportable) gestalts and widely distributed to the neocortex. Recent work has shown that concept cells, which are found in the medial temporal lobe of humans and activate when subjects perceive or think about specific concepts, fire when concepts are held in working memory [18].

Another theory that is relevant in the context of this discussion is Graziano's *attention schema theory* (AST) [19]. This theory posits that the brain constructs a simplified internal model, or "schema," of attention, which enables the brain to attribute awareness to itself, creating a self-referential understanding of attention. Graziano argues that consciousness arises from the brain's ability to model its own attentional processes. Although they are independent, it is possible to establish a connection between AST and GWT/GNW [19]. In the context of the attention schema theory, the global workspace serves as a platform for the brain to represent and manipulate the internal model of attention, contributing to the overall experience of consciousness.

4 A biologically inspired architecture for autotelic systems

Driven by evidence from biological systems and by the known behavior and limitations of existing models, we propose an architecture for intrinsically motivated systems that do open-ended learning. Our proposal is founded on three essential concepts. The first dictates the necessity of a two-system architecture, informed by CLS and GWT/GNW theories, as well as dual process theories [20, 21], which posit that there are two distinct information processing modes in the brain: one is fast, unconscious and high-bandwidth; the other is slow, deliberate and conscious. The second concept is that the behavior of the system is controlled by an attention schema [19], which takes into account available data to direct the focus of attention of the supervisor module. The third concept posits that the driving force behind a system capable of continuous learning is its inherent "desire" to enhance its understanding/model [22] and control [23] of the world, thereby sustaining goals that promote ongoing improvement in both the short and long term. This proposal combines ideas as diverse as

- *libido sciendi* ("passion for knowledge") [24], which provides the driver for learning;
- Friston's free energy principle for the brain [25] (which posits that the brain aligns its internal model of the world with the perceived external world by making predictions based on internal models of actions and updating them using sensory inputs), which provides the direction for learning;
- Popper's epistemological theory of falsificationism [26], whereby actions (experiments) are performed on the outside world to test "risky" hypotheses, *i.e.*, statements about the world of high generality, maximizing the rate of learning.

Figure 2 depicts the proposed *DeepThought* architecture, which includes several specific modules that play relevant roles. The **supervisor module**, at the top, enables short-term attention and memory and acts as the medium temporal lobe in the CLS theory, displaying fast adaptability, and controlling the global attention mechanism posited by GWT/GNW and AST. It receives inputs from the (system 1) language, vision, auditory, and world models at the bottom and focuses attention on specific inputs or on specific outputs of these models. As in GWT, tokens that deserve attention are broadcast to all the modules for explicit processing. The **deep reinforcement learning** (DRL) module uses the

4

inputs from the external world, under the control of the **attention schema**, and compares them with the internal world and self models, adjusting, as required, the model parameters.

As proposed by other authors [27], the DRL module is rewarded for action sequences that improve the predictions of the world model. These actions are suggested to the supervisor module, which may or may not decide to focus attention on them. The **language**, **auditory**, and **vision** modules, under the control of the attention schema, process information received from the outside world and/or from the supervisor (via the **embeddings store**). The embeddings in this store are derived by the joint encoding of external and internal multimodal token streams, as is done in CLIP [28].

The architecture also includes several memory components: a short-term **working memory**, which is part of the supervisory module, and a long-term episodic memory, which corresponds to the attention stream and is used to store explicit knowledge about the past. The succession of tokens that are the focus of attention is stored in **long-term memory** (the attention stream) for future retrieval and also to make meta-cognition possible. The system can describe the flow of reasoning, listing the series of sensations and actions that were the focus of attention. Semantic and procedural memory, on the other hand, are stored essentially in the language, vision, auditory, and world models, and can only be changed slowly, by adapting their parameters. We do not detail here the exact training mechanism that can be used to adjust the model parameters, mimicking the slow learning processes that take place in the neocortex, according to the CLS theory.
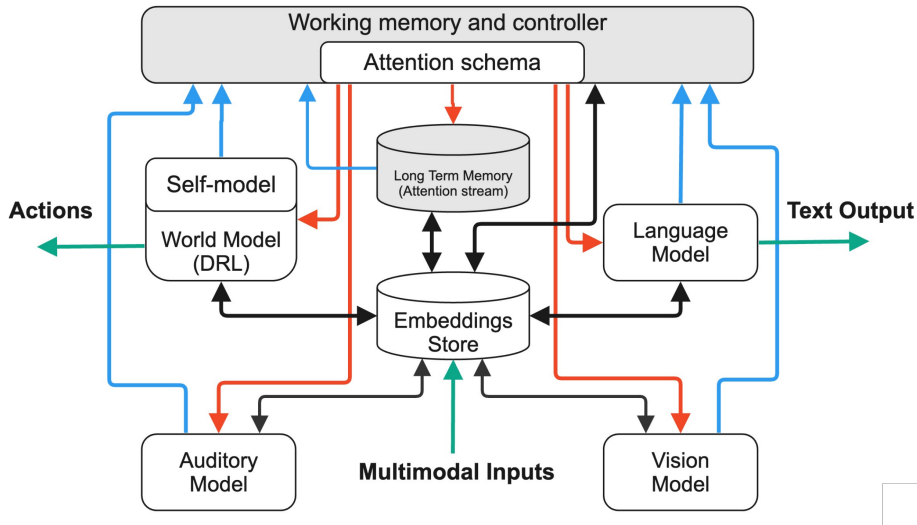

Figure 2: The DeepThought architecture: green arrows represent inputs and outputs, blue arrows represent module outputs, red arrows represent attention and control signals, and black arrows represent bidirectional interactions between modules and the embeddings store. The working memory and controller, which includes the attention schema, define the next inputs to the different modules.

5 Conclusions

Inspired by the current state of the art in cognitive language agents and by the GWT/GNW, AST, and CLS theories of cognition, we have proposed *DeepThought*, a two-system architecture for cognitive language agents that exhibit important characteristics of agency. The architecture includes several novel components, such as the integration of an attention schema, a clear definition of the role of the supervisory module, and the way system 1 and system 2 processes are combined.

Acknowledgements

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[4] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[5] Edward N. Zalta and Uri Nodelman, editors. *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Philosophy Department, Stanford University, 2012.

[6] John E Laird. *The SOAR cognitive architecture*. MIT Press, 2019.

[7] Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

[8] Mehmet Fırat and Saniye Kuleli. What if GPT4 became autonomous: The Auto-GPT project and use cases. *Journal of Emerging Computer Technologies*, 3(1):1 – 6, 2023.

[9] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[10] Cédric Colas, Laetitia Teodorescu, Pierre-Yves Oudeyer, Xingdi Yuan, and Marc-Alexandre Côté. Augmenting autotelic agents with large language models. *arXiv preprint arXiv:2305.12487*, 2023.

[11] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

[12] D Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 262(841):23–81, 1971.

[13] James L McClelland, Bruce L McNaughton, and Randall 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*, 102(3):419, 1995.

[14] James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974, 2020.

[15] Stanislas Dehaene, Michel Kerszberg, and Jean-Pierre Changeux. A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, 95(24):14529–14534, 1998.

[16] Bernard J Baars. The global workspace theory of consciousness. *The Blackwell companion to consciousness*, pages 236–246, 2007.

[17] Stanislas Dehaene. *Consciousness and the brain: Deciphering how the brain codes our thoughts*. Penguin, 2014.

[18] George A Mashour, Pieter Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene. Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5):776–798, 2020.

[19] Michael SA Graziano and Taylor W Webb. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*, 6:500, 2015.

[20] Keith Frankish. Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5(10):914–926, 2010.

[21] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[22] Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Anticipatory Behavior in Adaptive Learning Systems*, pages 48–74. Springer, 2008.

[23] Andrew G Barto. Intrinsic motivation and reinforcement learning. *Intrinsically motivated learning in natural and artificial systems*, pages 17–47, 2013.

[24] Volny Fages and Laurence Guignard. Introduction. libido sciendi. l'amour du savoir (1840-1900). *Revue d'histoire du XIXe siècle*, 2(57):9–20, 2018.

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

[26] Karl Popper. *Conjectures and refutations: The growth of scientific knowledge*. Routledge, 2014.

[27] Jürgen Schmidhuber. Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, 18(2):173–187, 2006.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.