Enabling High-Level Machine Reasoning with Cognitive Neuro-Symbolic Systems

Alessandro Oltramari

Bosch Center for Artificial Intelligence
2555 Smallman Street, Suite 302
Pittsburgh, Pennsylvania 15222 USA
alessandro.oltramari@us.bosch.com

Abstract

High-level reasoning can be defined as the capability to generalize over knowledge acquired via experience, and to exhibit robust behavior in novel situations. Such form of reasoning is a basic skill in humans, who seamlessly use it in a broad spectrum of tasks, from language communication to decision making in complex situations. When it manifests itself in understanding and manipulating the everyday world of objects and their interactions, we talk about *common sense* or *commonsense reasoning*. State-of-the-art AI systems don't possess such capability: for instance, Large Language Models have recently become popular by demonstrating remarkable fluency in conversing with humans, but they still make trivial mistakes when probed for commonsense competence; on a different level, performance degradation outside training data prevents self-driving vehicles to safely adapt to unseen scenarios, a serious and unsolved problem that limits the adoption of such technology. In this paper we propose to enable high-level reasoning in AI systems by integrating cognitive architectures with external neuro-symbolic components. We illustrate a hybrid framework centered on ACT−R, and we discuss the role of generative models in recent and future applications.

Introduction

A large part of neuro-symbolic systems is based on transforming symbolic knowledge into sub-symbolic representations that are suitable for infusion in data-driven learning algorithms: Knowledge Graph Embedding (KGE), among the others, is a prominent approach to reduce knowledge graph (KG) triples to latent vectors (Wang et al. 2017). Such transformation is instrumental to efficient computability of KG properties, as well as to application in a variety of downstream tasks: for instance, in (Wickramarachchi, Henson, and Sheth 2023) the authors leverage KGE methods to label unseen entities in autonomous driving datasets. Whether the KGE process is realized by geometric, tensor or deep learning models, the purpose is to *compress* KG structures into a low-dimensional space, where symbolic statements are replaced with dense, sub-symbolic expressions. Concatenation, non-linear mapping, attention-like mechanisms, gating mechanisms, are further methods to adapt knowledge structures to neural computations – e.g., (Peters et al. 2017; Strub et al. 2018; Margatina, Baziotis, and Potamianos 2019).

While knowledge-infusion can improve neural models, it is not sufficient to enable *high-level reasoning*, which is typically required by complex tasks such as natural language understanding, activity recognition, decision making in complex scenarios: latent, sub-symbolic expressions can only augment training signals with features derived from explicit semantic content, but this infusion process does neither carry any information about the reasoning mechanisms needed to process the learned knowledge, nor instruct the neural models on how those should unfold.

But, what do we mean with *high-level reasoning* and why is it important to endow artificial intelligent systems with such feature?

Problem Statement

We can define *high-level reasoning* as the capability to generalize over knowledge acquired via direct or mediated experience, and to exhibit robust behavior in novel situations. This definition is inspired by Kahneman's SYSTEM 2 mode of thought (Kahneman 2011). When *high-level reasoning* manifests itself in understanding and manipulating the everyday world of objects and their interactions, we talk about *common sense* or *commonsense reasoning*. State-of-the-art AI systems don't possess such capability: for instance, Large Language Models (LLMs) have recently become popular by demonstrating remarkable fluency in conversing with humans, but they still make trivial mistakes when probed for commonsense competence (see next section); on a different level, one of the motivations why the promise of autonomous cars hasn't panned out yet concerns performance degradation outside training data, which prevents self-driving vehicles to safely adapt to unseen scenarios.[1] Humans, on the opposite, are very good at generalizing from a few examples, and at filling the gaps in experience with reasoning: for in-

[1] A main weakness of deep learning approaches, as stated in a recent article (Bengio et al. 2019), is that 'current methods seem weak when they are required to generalize beyond the training distribution, which is what is often needed in practice', such as in safely maneuvering a vehicle.

stance, when asked about what happens after a bottle of red wine is thrown against a concrete wall, even children can answer with the utmost certainty that the bottle will shatter and the wall will be wet and red-stained – they can also easily infer that the impact between *any* fragile material and *any* hard surface *typically* ends with the former being substantially altered, if not destroyed; analogously, student drivers only need limited training to learn how to safely maneuver a car, adapting their knowledge and skills to novel situations. Compared to current AI systems based on GPU accelerated computing, human reasoning capabilities are impressive, even more so when we factor in what Herbert Simon used to call 'bounded rationality' (Simon 1955), i.e., the notion that human cognition operates with limited knowledge and is subject to time constraints – a heritage of evolution (Santos and Rosati 2015).

As these arguments suggest, a cognitive stance toward designing AI systems (Lieto 2021) seems to be key to enable high-level reasoning capabilities at the computational level: accordingly, we propose to complement cognitive architectures (Kotseruba and Tsotsos 2020; Langley, Laird, and Rogers 2009) with neuro-symbolic methods. In this paper we illustrate the blueprints of a *cognitive neuro-symbolic reasoning* system centered on the ACT-R[2] cognitive architecture (Anderson 1996), whose hybrid (symbolic and subsymbolic) mechanisms are well-suited for integration with neuro-symbolic algorithms and resources. Note that the proposed approach is applicable to any cognitive architecture whose properties are compatible with ACT-R, such as SOAR (Laird 2019) and SIGMA (Rosenbloom, Demski, and Ustun 2016): in fact, these three architectures have been grouped into the so-called 'Standard Model of the Mind' (Laird, Lebiere, and Rosenbloom 2017), an idea that has its roots in physics.[3] Note that the Standard Model of the Mind doesn't prescribe how to implement cognitively-inspired AI systems; rather, it aims to play the role of a conceptual framework of reference for developing them.

Motivations

Over the last decade, deep learning has yielded tremendous advancements in many AI fields, such as computer vision. For instance, neural models can achieve high accuracy in object detection when training and testing domains originate from the same data distribution. However, recent work shows that minimal/regional modifications implanted in the data at test time cause significant drop in accuracy (Eykholt et al. 2018; Rosenfeld, Zemel, and Tsotsos 2018). The examples documented in (Rosenfeld, Zemel, and Tsotsos 2018) are of particular interest, as they indicate how commonsense contextualization, by means of incorporating a priori structured knowledge into deep networks, can mitigate the effect of those perturbations, resulting in more robust performance (Marino, Salakhutdinov, and Gupta 2016). In general, a visual model suitably infused with knowledge extracted from semantic resources like CONCEPTNET (Speer, Chin, and Havasi 2017) can strengthen the connections holding within instances of the same conceptual domain (e.g., *couch*, *television*, *table*, *lamp* are located in living rooms) and discard out-of-context interpretations (e.g., no real *elephants* are located in living rooms, but photographs of elephant may be – figure 1 depicts such case).

When shifting to natural language, and to tasks like automated question answering, the key role played by knowledge-based contextualization for neural language models stands evident.[4] For instance, it has been demonstrated that using KG triples to disambiguate textual elements in a sentence, and embed the corresponding concepts and relations in neural language models (Devlin et al. 2018), significantly improves performance (Ma et al. 2021). In fact, despite of the impressive results that LLMs are producing in Natural Language Processing (Ma et al. 2019; Bauer and Bansal 2021; Shwartz et al. 2020), basic reasoning capabilities are still largely missing. This is also the reason why it's not appropriate to use 'Natural Language Understanding' to denote these tasks, because it would entail that robust and comprehensive reasoning capabilities are present (McShane 2017). Let's expand on this argument and consider a few representative examples.

In ProtoQA (Boratko et al. 2020), GPT-2 (Dale 2021) fails to select options like 'pumpkin', 'cauliflower', 'cabbage' as top candidates, for the question 'one vegetable that is about as big as your head is?': instead, 'broccoli', 'cucumber', 'beet', 'carrot' are predicted. In this case, the different models learn some essential properties of vegetables from the training data, but do not seem to acquire the capability of comparing their size to that of other types of objects, revealing a substantial lack of *analogical reasoning* (Ushio et al. 2021). The same issues are observed when CHATGPT, a recent popular version of GPT-3 optimized for conversations, is considered: the main difference is that CHATGPT is capable of generating plausible answers when the question is submitted literally, but often fails to do so when the verbal expression 'about as big as' is paraphrased with alternative forms like 'about the same size', 'about the same shape', 'comparable to', etc. This 'hypersensitivity' to surface-level linguistic features – an epiphenomenon of the model's incapability to generalize over textual variations of the same content – seem to indicate that the model cannot perform the necessary (analogical) reasoning steps needed to correctly answer to the question. Along these lines, recent work (Ettinger 2020) has shown that lack of complex inferences, role-based event prediction, and understanding the conceptual impact of negation, are some of the weaknesses diagnosed when BERT (Devlin et al. 2018), one of promi-

[2]Abbreviation of 'Adaptive Control of Thought, Rational'.

[3]For a brief introduction to the *Standard Model of Particle Physics*, see this resource from the U.S. Department of Energy: https://www.energy.gov/science/doe-explainsthe-standard-model-particle-physics

[4]We use 'language model', 'neural language model' and 'large language model' as interchangeable terms, as they commonly refer to the same neural architecture based on multi-headed self-attention mechanisms (Vaswani et al. 2017); however, computational power significantly differs as function of the specific implementations (e.g., BERT has 6 blocks with 12 heads, GPT-3 has 24 blocks and 48 heads), and of the size of training datasets (CHATGPT has been trained on a massive corpus – 570 GB – of text data).

Figure 1: The *Elephant in the Room*: the probability that a label assigned by an object detection system is correct increases when the context is factored in: in this example, the label 'elephant' could plausibly denote a picture of the pachyderm, but not the pachyderm itself.

nent open source language models, is applied to benchmark datasets. ProtoQA again provides good examples of these deficiencies: in general, neural models struggle to correctly interpret the scope of modifiers like 'not' (*reasoning under negation*), 'often' and 'seldom' (*temporal reasoning*). Regarding the latter, in task 14 of bAbI (Weston et al. 2015), a comprehensive benchmark challenge designed by Facebook Research, neural language systems exhibit variable accuracy in grasping temporal ordering entailed by prepositions like 'before' and 'after'. Similarly, in bAbI task 17, which concerns *spatial reasoning*, LLM-based systems fail to infer basic positional information that require interpreting the semantics of 'to the left/right of', 'above/below', etc. If such systems are inaccurate when dealing with common characteristics of the physical world, their performance doesn't improve when sentiments are considered: for instance, in SocialIQA (Sap et al. 2019), given a context like 'in the school play, Robin played a hero in the struggle to death with the angry villain', models are unable to consistently select 'hopeful that Robin will succeed' over 'sorry for the villain' when required to pick the correct answer to 'how would others feel afterwards?'. It's not surprising that *reasoning about emotional reactions* represents a difficult task for pure learning systems, when we consider that such form of inference is deeply rooted in the sphere of human experiences and social life, which involves a 'layered' understanding of mental attitudes, intentions, motivations, emotions, and of the events that trigger them.

The qualitative analysis presented above suggests that neural models struggle to perform well in tasks that require high-level reasoning. But, are neuro-symbolic approaches sufficient to overcome these limitation? Latent expressions can augment training signals with sub-symbolic features derived from explicit semantic content, but knowledge infusion *per se* doesn't determine how inference processes are conducted. Relevant work in this space shows how deep neural models can replicate logical reasoning (Ebrahimi, Eberhart, and Hitzler 2021; Garcez et al. 2022), but it doesn't follow that any form of logical reasoning that is provably reducible to learning algorithms, should also be systematically reduced to it – this would be a requirement only for tightly-coupled neuro-symbolic systems (Kautz 2022; Garcez and Lamb 2023).

Accordingly, in the next section we make the case for developing an AI framework where the ACT−R architecture is *loosely-coupled* with neuro-symbolic components, to enable high-level reasoning.

Method

Cognitive architectures attempt to capture at the computational level the invariant mechanisms of human cognition, including those underlying the functions of control, learning, memory, adaptivity, perception and action. ACT-R (Anderson 1996), in particular, is designed as a hybrid modular framework including perceptual, motor and memory components, synchronized by a procedural module through limited capacity buffers. Over the years, ACT-R has accounted for a broad range of tasks at a high level of fidelity, reproducing aspects of complex human behavior, from everyday activities like event planning (Somers, Oltramari, and Lebiere 2020) and car driving (Cina and Rad 2023), to highly technical tasks such as piloting an airplane (Chen et al. 2021), and monitoring a network to prevent cyber-attacks (Ben-Asher et al. 2015). ACT-R has been used as a component in pipelines that include learning algorithms (e.g., biologically-inspired neural networks (Jilk et al. 2008)) and external semantic resources (e.g., (Oltramari and Lebiere 2012; Emond 2006)): along this line of research, we claim that integrating ACT-R – or any compatible cognitive architecture – with neuro-symbolic components is instrumental to enable high-level machine reasoning.

Figure 2 provides a compact visualization of our proposed framework: the boxes in blue, enclosed in the grey rectangle, represent the default components of ACT−R, those in green the neuro-symbolic extensions.

The integration would occur along three main directions:

- **knowledge** $\leftrightsquigarrow$ **memory**: the external symbolic module, which can include background/domain knowledge graphs (KG), lexical resources (LR), rule bases (RB), and a suitable inference engine, is linked to the declarative

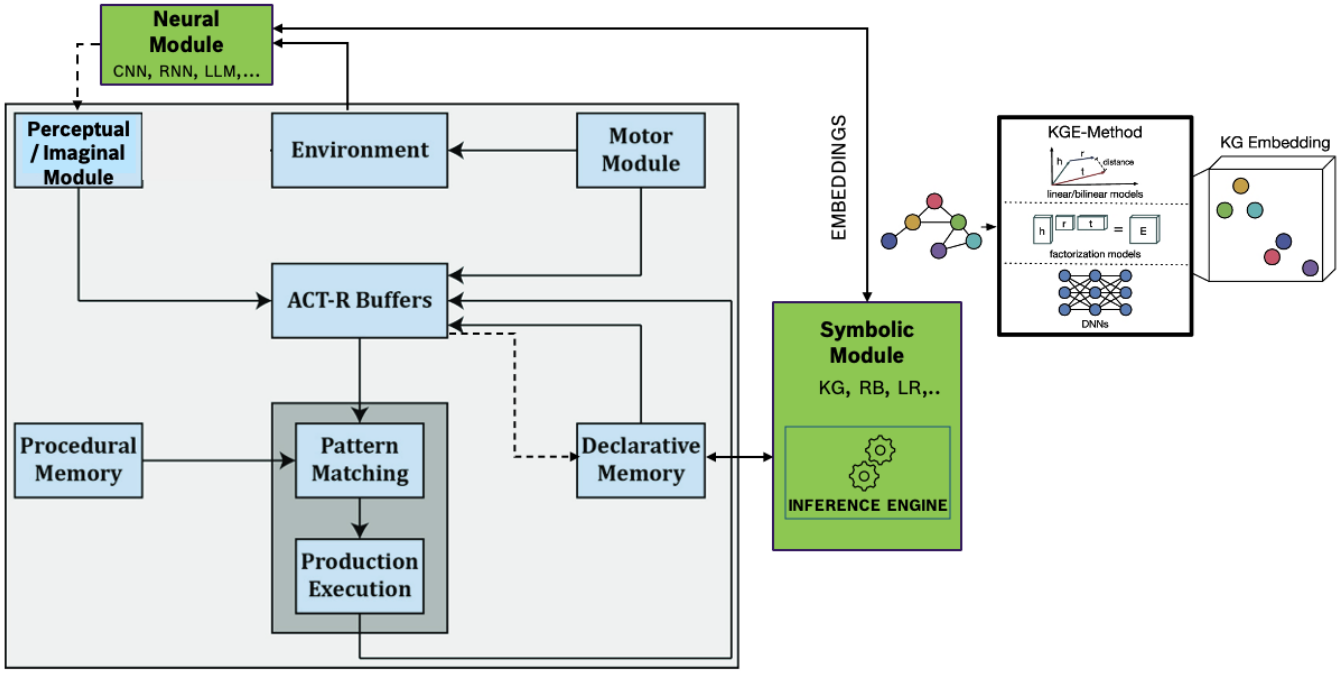


Figure 2: ACT-R integrated with neuro-symbolic modules.

memory. This is a two-way integration: the symbolic module can be *read* or *written* by ACT−R, where the latter operation is triggered when populating or pruning world knowledge is needed as part of task execution.

- **neural** $\rightsquigarrow$ **perception**: the neural module, which can include convolutional, recurrent, long-short-term memory networks, generative models, etc., is trained, fine-tuned, or prompted with data processed from the environment, providing relevant patterns of information to the perceptual or imaginal module. This integration bypasses the direct connection holding – in standard ACT−R – between the perceptual module and the environment.[5]
- **knowledge** $\rightsquigarrow$ **neural**: adequately-selected embedding mechanisms govern knowledge-infusion in the neural module, enabling knowledge-based contextualization of patterns of information distilled from the environment, which are subsequently channeled into ACT−R buffers.

If the mutual connections between the two intertwined neuro-symbolic modules and ACT-R can be used to combine rich semantic contents with scalable learning functionalities, they don't *per se* bring about high-level reasoning: this capability also requires two features of the integrated framework, namely the cognitive architecture's own procedural module and a proper inference engine in the external symbolic module.

The procedural module matches the content of the other module buffers and coordinates their activity using production rules, which are 'condition-action' pairs tied to the task at hand. Productions use an utility-based computation to select, from a set of task-specific plausible rules, the single rule that is executed at any point in time. For instance, when building a recommendation system to support a mechanic in troubleshooting a car engine, a relevant situation that needs to be covered is a vehicle that doesn't start but has power; in this example, a high-utility production rule should capture the following heuristic: *if the engine holds compression well, and the fuel system is working correctly, then the spark plugs should be checked*. The variables in these rule conditions would need to be filled with actual empirical observations and measurements, as it is often the case when cognitive architectures are applied in real-world scenarios: in our example, such evidence could be actually gathered by a real technician using the recommendation system in a human-machine-teaming fashion, a type of approach that falls under the 'cognitive model as oracle' paradigm (Lebiere et al. 2022).

The inference engine in the symbolic module is used to derive knowledge from assertions in the semantic resource of reference, a well-known feature of symbolic AI systems. What is important to stress here, is that – in our proposal – this form of logic-based reasoning would realize two functions: 1) provide a combination of asserted and inferred knowledge that ACT-R declarative memory can process and pass to the production system; 2) support knowledge-infusion into neural modules. The first functionality would help to decouple basic forms of reasoning, e.g. temporal and spatial[6] , from cognitive assessments performed by the pro-

[5]Such connection assumes symbolic representations of visual and auditory signals being available to the architecture through pre-processing.

[6]E.g., Region-Connection-Calculus (Cohn et al. 1997) for spa-

duction system on conditional actions. Such feature makes our proposed system efficient, as ACT-R productions are not well-suited for logical reasoning. The second functionality would allow pre-training, fine-tuning, or prompting a neural-model on both asserted and inferred knowledge: this can provide ACT-R perceptual model with more informative patterns than just those obtained by processing raw data.

It's worth making a final consideration here: the framework introduced in this section is complemental to the body of work that investigates how neuro-symbolic systems can be leveraged to realize human-like cognitive reasoning (Garcez, Lamb, and Gabbay 2008): in our proposal, ACT-R is interfaced with neuro-symbolic components, whereas – in the approaches reviewed by Garcez et al. – neuro-symbolic frameworks are used to solve cognitive tasks. The difference lies on whether cognitive processes are considered *first class citizens* or not.

Discussion: The Role of Generative AI in Cognitive Neuro-Symbolic Reasoning

As seen in the previous section, our proposed framework doesn't require or commit on a specific neural architecture. However, generative AI, and specifically large language models, will play an increasingly relevant role in enabling high-level reasoning based on *cognitive neuro-symbolic systems*. In the next two sections we will briefly outline present research in this field, and sketch what we think are promising developments.

Related Work

The importance of integrating cognitive mechanisms into data-driven AI systems has been recently acknowledged by one of the key figures in deep learning, Yann LeCun: in a position paper published in 2022 (LeCun 2022), he described a biologically-inspired cognitive architecture, where a so-called *configurator* orchestrates information provided by different modules, such as the *perception module* and the *world model module*, which replicate the functions emerging from prefrontal-cortical processes. Furthermore, a *motivation model* – designed to mimic the role of the amygdala in producing basic emotional states like pain and pleasure – is used to compute intrinsic costs associated with current and future actions, a mechanism that is instrumental to inform predictive capabilities. It's relevant to point out that there has been extensive research on mapping cognitive architectures to brain areas/processes – e.g., (Borst et al. 2015) – and that an established scientific community has been working on biologically-inspired approaches to cognitive architectures since the early 2000's (the BICA international conference has reached its 14th edition[7]).

In line with the current trend of investigating computational models of cognition in the context of large-scale neural networks, a recent blog (Weng 2023) provides an overview of how LLMs could be used to control autonomous agents. It goes beyond the scope of our contribution to review in detail the papers mentioned in the blog, but it's beneficial to highlight some of the most interesting topics.

In (Wei et al. 2022) the authors leverage *chain-of-thought* prompting with PALM 540B (Chowdhery et al. 2022) for task-decomposition: despite of their reported success, using prompting to generate fine-grained reasoning steps does not always yield consistent results, as shown by (Chen, Zaharia, and Zou 2023) for different versions of GPT-4 (OpenAI 2023). The same work also indicates that, even when reasoning steps are correctly reproduced, they don't always match with the model selecting the correct solution/answer to a problem/question. Another paper surveyed in the blog (Park et al. 2023) focuses on using GPT-4 to build *believable agents* for a sandbox environment [8]. According to the authors, cognitive architectures would not have the same degree of flexibility (and scalability) that modern generative models provide when building AI agents, as the former depend on hand-crafting rules, thus applicable only to narrow, closed-world contexts. However, this is a partial account of the state of the art: for instance, production compilation, ACT-R's rule learning mechanism, allows to learn new, task-specific production rules that directly implement the relevant action(s) for a particular state (Taatgen, Huss, and Anderson 2006); moreover, to assess which stimuli from an environment are relevant for an agent to act upon, researchers have developed mechanisms like instance-based learning, a type of reinforcement learning (Sutton and Barto 2018), which can be plugged into ACT-R (Gonzalez, Lerch, and Lebiere 2003). One may also question the claim on generative models' flexibility: in fact, the scope of such capability is not the real world, with its ever-changing situations, but rather some emerging patterns in the text-based training data, which are biased interpretations of the real world. Incidentally, this lack of 'grounding' is also at the origin of LLMs' widely-documented hallucination problem – for an introduction to this phenomenon, see (Ji et al. 2023).[9]

Future Work

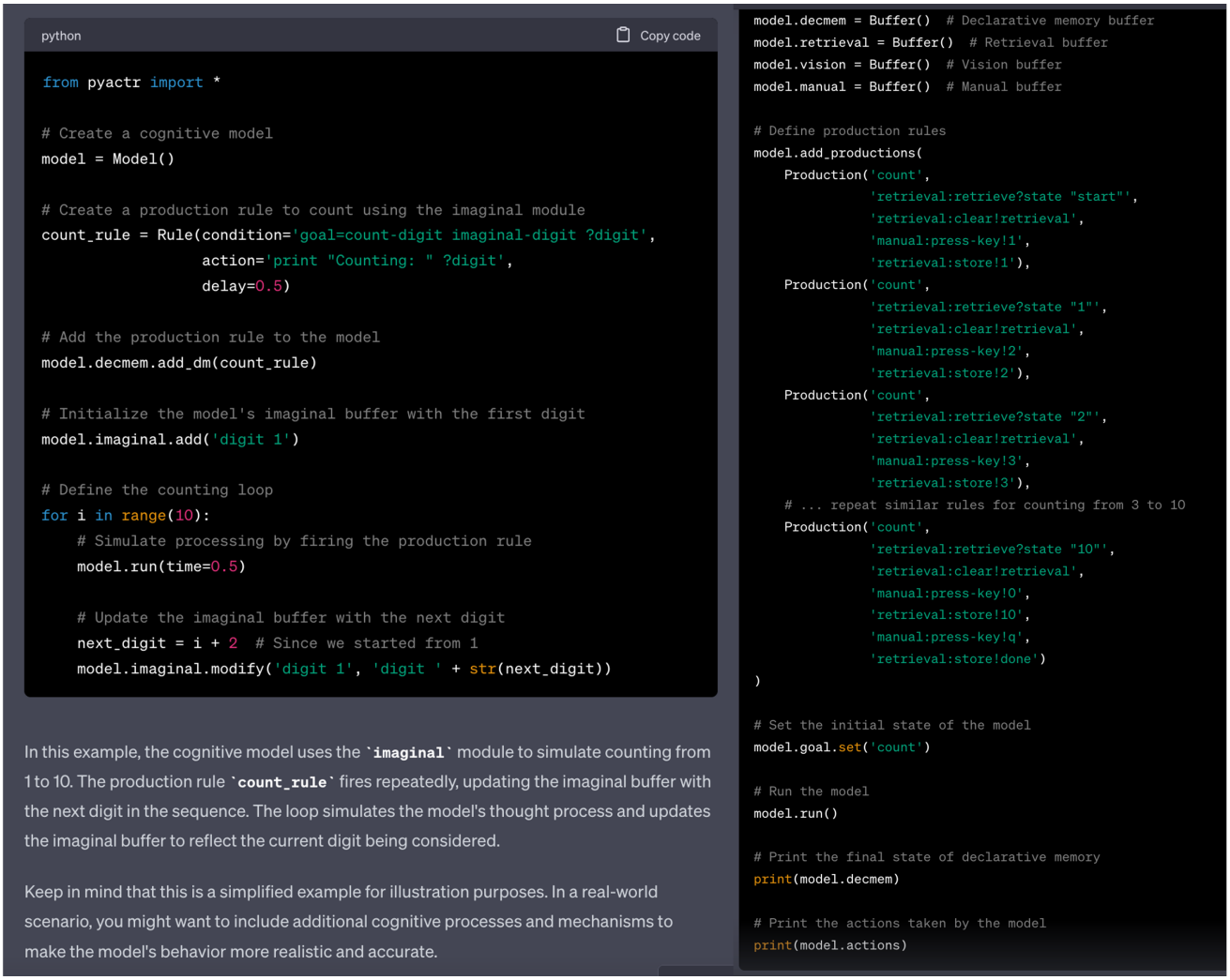
As the overview in the previous section suggests, there are intrinsic limitations in utilizing a LLM as *orchestrator* for intelligent agents. In this regard, we can distill two main reasons for selecting a cognitive architecture over a LLM: a) the inner functioning of the former is transparent, whereas the latter is a 'black-box' (Castelvecchi 2016); b) the former is designed to replicate the invariant mechanisms of human cognition, the latter is engineered to produce human-grade linguistic *behavior*, which cognitive properties can only be ascribed to. By and large, what the state-of-the-art suggests is that a synergistic integration of these cognitive architectures and LLMs can help to maximize their relative strengths

tial reasoning, Allen's axioms for temporal reasoning (Allen and Ferguson 1994).

[7]See: https://bica2023.org/cfp/

[8]Inspired by the video-game 'The Sims': https://www.ea.com/games/the-sims

[9]There is an interesting analogy between Plato's Cave myth (Jowett et al. 1873), where shadows projected on a blank wall were all that prisoners could use to understand reality, thus misinterpreting it, and LLMs's generating inaccurate statements about the world, based on biased data patterns.



python

Copy code

```python
from pyactr import *

# Create a cognitive model
model = Model()

# Create a production rule to count using the imaginal module
count_rule = Rule(condition='goal=count-digit imaginal-digit ?digit',
                  action='print "Counting: " ?digit',
                  delay=0.5)

# Add the production rule to the model
model.decmem.add_dm(count_rule)

# Initialize the model's imaginal buffer with the first digit
model.imaginal.add('digit 1')

# Define the counting loop
for i in range(10):
    # Simulate processing by firing the production rule
    model.run(time=0.5)

    # Update the imaginal buffer with the next digit
    next_digit = i + 2  # Since we started from 1
    model.imaginal.modify('digit 1', 'digit ' + str(next_digit))
```

In this example, the cognitive model uses the `imaginal` module to simulate counting from 1 to 10. The production rule `count_rule` fires repeatedly, updating the imaginal buffer with the next digit in the sequence. The loop simulates the model's thought process and updates the imaginal buffer to reflect the current digit being considered.

Keep in mind that this is a simplified example for illustration purposes. In a real-world scenario, you might want to include additional cognitive processes and mechanisms to make the model's behavior more realistic and accurate.

```python
model.decmem = Buffer()  # Declarative memory buffer
model.retrieval = Buffer()  # Retrieval buffer
model.vision = Buffer()  # Vision buffer
model.manual = Buffer()  # Manual buffer

# Define production rules
model.add_productions(
    Production('count',
               'retrieval:retrieve?state "start"',
               'retrieval:clear!retrieval',
               'manual:press-key!1',
               'retrieval:store!1'),
    Production('count',
               'retrieval:retrieve?state "1"',
               'retrieval:clear!retrieval',
               'manual:press-key!2',
               'retrieval:store!2'),
    Production('count',
               'retrieval:retrieve?state "2"',
               'retrieval:clear!retrieval',
               'manual:press-key!3',
               'retrieval:store!3'),
    # ... repeat similar rules for counting from 3 to 10
    Production('count',
               'retrieval:retrieve?state "10"',
               'retrieval:clear!retrieval',
               'manual:press-key!0',
               'retrieval:store!10',
               'manual:press-key!q',
               'retrieval:store!done')
)

# Set the initial state of the model
model.goal.set('count')

# Run the model
model.run()

# Print the final state of declarative memory
print(model.decmem)

# Print the actions taken by the model
print(model.actions)
```

Figure 3: Without adequate instructions, CHAT-GPT opts for a compact model of counting from 1 to 10, which is not substantially distinct from a simple `for` loop (left-side). Interestingly, the chatbot suggests to *include additional cognitive processes and mechanisms to make the model's behavior more realistic and accurate* (bottom-left): this is what actually happens when the LLM is instructed to use all ACT-R modules (right-side). Far from being exhaustive, this example provides some evidence of the feasibility of scaling cognitive models via LLMs.

and mitigate their weaknesses, fostering the creation of more advanced AI systems, capable of high-level reasoning. In particular, (1) *scaling cognitive models via* LLMs and (2) *prompt-engineering* LLM*s with cognitive models* can be seen as novel approaches in this direction; they would actually be complemental, as (1) is a method to automatize the creation of cognitive models using generative AI, whereas (2) is a method to ground generative AI on computational artifacts that reflect principled cognitive theories.

1. ***Scaling cognitive models via* LLMs**. A cognitive architecture is a generic framework to develop cognitive models, which are, conversely, tied to specific tasks and domains: the process of developing cognitive models is still largely manual, and thus affected by lack of scalability. Because LLMs have proven to be effective in generating code across a variety of programming languages (Gozalo-Brizuela and Garrido-Merchan 2023), they could also be leveraged to produce software implementations of cognitive models. Initial experiments performed by asking CHAT-GPT to generate basic cognitive models using a novel library, i.e., `PyACT-R`[10], show that the OpenAI's signature LLM learns to correctly generate compact Python snippets, although it only makes marginal use of ACT-R modules and buffers. In order to achieve such level of sophistication in cognitive model design, CHAT-GPT needs to be prompted with relevant instructions about which mechanisms of a cognitive architecture it should use (see figure 3).
2. ***Prompt-engineering* LLM*s* with cognitive models**. Using LLMs in domain-specific applications requires either fine-tuning on a target dataset, or prompt-engineering with adequate contextual knowledge. In many use cases, well-curated data are either unavailable or too time-consuming to collect at scale, making the latter more convenient and efficient. When the goal is to turn a LLM into a reliable decision support system, the 'grounding' problem mentioned earlier also extends to the cognitive dimension: that is, such system would need to be based on shared interpretations of reality as well as on sound reasoning steps, from a cognitive-decisional standpoint. In fact, it'd be difficult to conceive such a system as trustworthy if hallucinations on both factual knowledge and on inferential mechanisms were widespread. To this end, prompting a LLM with key steps of a cognitive model's reasoning process, the so-called *trace*, would be instrumental to mitigate the second type of hallucinations. Such steps *de facto* represent the introspective stages of a cognitive model, and of a cognitive neuro-symbolic reasoning system based on it.

Conclusion

In the current debate on the limits of deep neural networks, the split is oftentimes between those who think that *more data* is the panacea, and those who support designing systems that integrate learning approaches with other processing elements, such as knowledge representation and reasoning, statistical algorithms, human-in-the-loop methods. In this paper, which echoes the second category, we made the case for adopting a cognitive approach to perform that integration, inspired by the results that architectures like ACT-R have produced, over the last decades, in replicating complex human tasks at the machine level. We described the main components of a *cognitive neuro-symbolic reasoning system*, outlined their respective functionalities, and discussed related and future work in the area of generative AI.
At the end, we don't assume or prove that using cognitive architectures is the only possibility to equip machines with high-level, human-like reasoning: however, to paraphrase (Mittal, Bengio, and Lajoie 2022),through a diversity of scientific explorations, we'll increase our chances to find the ingredients we are missing.

[10]https://github.com/jakdot/pyactr

References

Allen, J. F.; and Ferguson, G. 1994. Actions and events in interval temporal logic. *Journal of logic and computation*, 4(5): 531–579.

Anderson, J. R. 1996. ACT: A simple theory of complex cognition. *American psychologist*, 51(4): 355.

Bauer, L.; and Bansal, M. 2021. Identify, Align, and Integrate: Matching Knowledge Graphs to Commonsense Reasoning Tasks. In *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2259–2272. Online: Association for Computational Linguistics.

Ben-Asher, N.; Oltramari, A.; Erbacher, R. F.; and Gonzalez, C. 2015. Ontology-based Adaptive Systems of Cyber Defense. In *STIDS*, 34–41.

Bengio, Y.; Deleu, T.; Rahaman, N.; Ke, R.; Lachapelle, S.; Bilaniuk, O.; Goyal, A.; and Pal, C. 2019. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*.

Boratko, M.; Li, X. L.; Das, R.; O'Gorman, T.; Le, D.; and McCallum, A. 2020. ProtoQA: A Question Answering Dataset for Prototypical Common-Sense Reasoning. *arXiv preprint arXiv:2005.00771*.

Borst, J. P.; Nijboer, M.; Taatgen, N. A.; van Rijn, H.; and Anderson, J. R. 2015. Using data-driven model-brain mappings to constrain formal models of cognition. *PLoS One*, 10(3): e0119673.

Castelvecchi, D. 2016. Can we open the black box of AI? *Nature News*, 538(7623): 20.

Chen, H.; Liu, S.; Pang, L.; Wanyan, X.; and Fang, Y. 2021. Developing an improved ACT-R model for pilot situation awareness measurement. *IEEE Access*, 9: 122113–122124.

Chen, L.; Zaharia, M.; and Zou, J. 2023. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Cina, M.; and Rad, A. B. 2023. Categorized review of drive simulators and driver behavior analysis focusing on ACT-R architecture in autonomous vehicles. *Sustainable Energy Technologies and Assessments*, 56: 103044.

Cohn, A. G.; Bennett, B.; Gooday, J.; and Gotts, N. M. 1997. Qualitative spatial representation and reasoning with the region connection calculus. *geoinformatica*, 1: 275–316.

Dale, R. 2021. GPT-3: What's it good for? *Natural Language Engineering*, 27(1): 113–118.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ebrahimi, M.; Eberhart, A.; and Hitzler, P. 2021. On the Capabilities of Pointer Networks for Deep Deductive Reasoning. *arXiv preprint arXiv:2106.09225*.

Emond, B. 2006. WN-LEXICAL: An ACT-R module built from the WordNet lexical database. In *Proceedings of the Seventh International Conference on Cognitive Modeling*, 359–360.

Ettinger, A. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8: 34–48.

Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; and Song, D. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1625–1634.

Garcez, A. d.; Bader, S.; Bowman, H.; Lamb, L. C.; de Penning, L.; Illuminoo, B.; Poon, H.; and Gerson Zaverucha, C. 2022. Neural-symbolic learning and reasoning: A survey and interpretation. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342: 1.

Garcez, A. d.; and Lamb, L. C. 2023. Neurosymbolic AI: The 3 rd wave. *Artificial Intelligence Review*, 1–20.

Garcez, A. S.; Lamb, L. C.; and Gabbay, D. M. 2008. *Neural-symbolic cognitive reasoning*. Springer Science & Business Media.

Gonzalez, C.; Lerch, J. F.; and Lebiere, C. 2003. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4): 591–635.

Gozalo-Brizuela, R.; and Garrido-Merchan, E. C. 2023. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. *arXiv preprint arXiv:2301.04655*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Jilk, D. J.; Lebiere, C.; O'Reilly, R. C.; and Anderson, J. R. 2008. SAL: An explicitly pluralistic cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 20(3): 197–218.

Jowett, B.; et al. 1873. *The dialogues of Plato*, volume 4. Scribner, Armstrong.

Kahneman, D. 2011. *Thinking, fast and slow*. macmillan.

Kautz, H. 2022. The third ai summer: Aaai robert s. engelmore memorial lecture. *AI Magazine*, 43(1): 105–125.

Kotseruba, I.; and Tsotsos, J. K. 2020. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 53(1): 17–94.

Laird, J. E. 2019. *The Soar cognitive architecture*. MIT press.

Laird, J. E.; Lebiere, C.; and Rosenbloom, P. 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.

Langley, P.; Laird, J. E.; and Rogers, S. 2009. Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2): 141–160.

Lebiere, C.; Cranford, E.; Martin, M.; Morrison, D.; and Stocco, A. 2022. Cognitive architectures and their applications. In *Proceedings of IEEE CIC*.

LeCun, Y. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62.

Lieto, A. 2021. *Cognitive design for artificial minds*. Routledge.

Ma, K.; Francis, J.; Lu, Q.; Nyberg, E.; and Oltramari, A. 2019. Towards Generalizable Neuro-Symbolic Systems for Commonsense Question Answering. In *Proc. of the First Workshop on Commonsense Inference in Natural Language Processing*, 22–32.

Ma, K.; Ilievski, F.; Francis, J.; Bisk, Y.; Nyberg, E.; and Oltramari, A. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Proc. of 35th AAAI Conference on Artificial Intelligence*.

Margatina, K.; Baziotis, C.; and Potamianos, A. 2019. Attention-based conditioning methods for external knowledge integration. *arXiv preprint arXiv:1906.03674*.

Marino, K.; Salakhutdinov, R.; and Gupta, A. 2016. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*.

McShane, M. 2017. Natural language understanding (NLU, not NLP) in cognitive systems. *AI Magazine*, 38(4): 43–56.

Mittal, S.; Bengio, Y.; and Lajoie, G. 2022. Is a modular architecture enough? *Advances in Neural Information Processing Systems*, 35: 28747–28760.

Oltramari, A.; and Lebiere, C. 2012. Using ontologies in a cognitive-grounded system: automatic action recognition in video surveillance. In *Proceedings of the 7th International Conference on Semantic Technology for Intelligence, Defense, and Security*. Citeseer.

OpenAI, R. 2023. GPT-4 technical report. *arXiv*, 2303–08774.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Peters, M. E.; Ammar, W.; Bhagavatula, C.; and Power, R. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Rosenbloom, P. S.; Demski, A.; and Ustun, V. 2016. The Sigma cognitive architecture and system: Towards functionally elegant grand unification. *Journal of Artificial General Intelligence*, 7(1): 1.

Rosenfeld, A.; Zemel, R.; and Tsotsos, J. K. 2018. The elephant in the room. *arXiv preprint arXiv:1808.03305*.

Santos, L. R.; and Rosati, A. G. 2015. The evolutionary roots of human decision making. *Annual review of psychology*, 66: 321–347.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proc. of EMNLP-IJCNLP*, 4463–4473.

Shwartz, V.; West, P.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2020. Unsupervised Commonsense Question Answering with Self-Talk. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4615–4629. Online: Association for Computational Linguistics.

Simon, H. A. 1955. A behavioral model of rational choice. *The quarterly journal of economics*, 99–118.

Somers, S.; Oltramari, A.; and Lebiere, C. 2020. Cognitive Twin: A Cognitive Approach to Personalized Assistants. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1)*.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Strub, F.; Seurin, M.; Perez, E.; De Vries, H.; Mary, J.; Preux, P.; and Pietquin, A. C. 2018. Visual reasoning with multi-hop feature modulation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 784–800.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Taatgen, N. A.; Huss, D.; and Anderson, J. R. 2006. How cognitive models can inform the design of instructions. In *Proceedings of the seventh international conference on cognitive modeling*, 304–309. Citeseer.

Ushio, A.; Espinosa-Anke, L.; Schockaert, S.; and Camacho-Collados, J. 2021. BERT is to NLP what AlexNet is to CV: Can Pre-Trained Language Models Identify Analogies? *arXiv preprint arXiv:2105.04949*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Q.; Mao, Z.; Wang, B.; and Guo, L. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12): 2724–2743.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Weng, L. 2023. LLM-powered Autonomous Agents. *lilianweng.github.io*.

Weston, J.; Bordes, A.; Chopra, S.; Rush, A. M.; van Merriënboer, B.; Joulin, A.; and Mikolov, T. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Wickramarachchi, R.; Henson, C.; and Sheth, A. 2023. CLUE-AD: A Context-Based Method for Labeling Unobserved Entities in Autonomous Driving Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 16491–16493.