Metacognition-Driven Preprocessing for Optimized Artificial Intelligence Performance

Naavya Shetty
Undergraduate in Computer Science and Philosophy
Department of Philosophy
University of Illinois Urbana-Champaign
Illinois, United States
e-mail: leltuz404@gmail.com

***Abstract*—Machine cognition is currently heavily speed-based. Directly tackling inputs with computation often leads to inefficient steps, such as performing redundant or repetitive computation, or execution without assessing whether a task is within computational capacity. This paper proposes a preprocessing metacognitive system to be implemented in a manner such that it screens all input requests, creating a strategic 'bottleneck' to filter, redirect or halt the flow of control before computation begins. The findings theorise improved accuracy, reliability and resource-management, strengthening the argument for making metacognition an essential component of artificial intelligence.**

Keywords-artificial intelligence; resource optimization; self-modifying machines.

I. Introduction

Formulating as accurate a response in as little time as possible is the goal computation strives to achieve, but there is no system in place to determine whether it has the computational ability to do so, nor to overcome redundant operations, thus failing to optimize processing efficiency. Implementing solutions to these issues requires a kind of 'awareness' in computation, which poses challenges in terms of defining self-assessment standards, developing algorithms to monitor computational efficiency, and integrating self-adapting decision making processes. Machines lack the concept of cognitive overload, making it difficult to ascertain when an operation should be adjusted or entirely terminated before execution. There also exists the issue of balancing computational overhead with the benefits of self-regulation.

Existing research in this space such as Schaeffer [1] primarily focuses on detecting suboptimal actions in various forms of machine learning contexts. The following research, however, aims to extend this by developing a preprocessing metacognitive system that not only identifies inefficiencies but also proactively consults a database of prior experiences to optimize computational resources. This broader scope addresses not just action evaluation but also strategic planning and resource management, offering a more comprehensive solution to the efficiency of artificial intelligence systems.

The need for a solution to this problem lies in the lack of computational efficiency leading to wasted resources, increased latency, and suboptimal decision-making, which could lead to cascading inefficiencies or altogether failure, especially in high-stakes applications such as autonomous systems and large-scale simulations. Addressing this gap by introducing a metacognition-based system could enable the existing infrastructure to allocate resources more dynamically, recognize when to rethink strategies, and improve performance while simultaneously minimizing computation.

The core scientific problem is the absence of metacognition in current artificial intelligence software, preventing systems from evaluating their computational strategies and optimizing efficiency. Unlike human cognition, which involves self-regulation and resource allocation, current systems lack the mechanism to assess when an operation is redundant or inefficient. The challenge lies in developing an architecture that enables such a system to recognize its own performance constraints and adjust computational processes without excessive overhead.

Through this paper, the definition of metacognition will be analyzed and mapped into machine cognition in such a way that it sheds light on and provides impetus to a possible adjacent feature of artificial general intelligence that can enable evaluation of its own computational limits, streamline decision-making, and reduce inefficiencies in problem-solving. What is sought here is a form of automative 'wisdom' over intelligence. In other words, the goal is to augment and improve decision-making abilities in machines, in addition to being able to deliver intelligent responses.

Some questions that will be tackled through this paper are how metacognition can be modeled and implemented in artifical intelligence systems to enhance preprocessing strategies that can enable self-evaluation of computational efficiency, what mechanisms allow machines to adjust preprocessing strategies dynamically, how machines can detect inefficiencies or redundant operations in data preprocessing, and how metacognitive preprocessing improves the adaptability and robustness of current artificial intelligence systems.

The purpose of this paper is to explore how metacognitive principles can be integrated into artificial systems to enhance their decision-making efficiency. By equipping such a system with a form of self-awareness regarding its computational processes, the research aims to reduce redundant operations, optimize resource allocation, and improve overall system intelligence. This work lays the foundation for models that not only generate intelligent responses but also assess and refine their reasoning processes.

This proposal is not without its limitations. Implementing metacognition in computation introduces computational overhead, which could paradoxically reduce efficiency if not carefully managed. Additionally, defining an objective metric for 'effort' in machine cognition remains an open challenge.

While biological systems can optimize through evolutionary processes, artificial systems lack intrinsic motivation, making it difficult to determine when computational adjustments are necessary. Furthermore, existing architectures may require fundamental modifications to accommodate self-regulation mechanisms effectively.

This paper first delves into the background of artificial intelligence systems and metacognition, as well as some relevant definitions for the proposed theory. It then describes the proposed theory, describing the features of such a theory, the requirements for building it, and highlights necessary qualities in the implementation. Finally, the paper discusses some rebuttals that may be raised and provides an answer to them.

II. Background

Current endeavors towards building artificial intelligence are geared towards creating systems that can emulate and surpass human intelligence. The Dartmouth meeting of 1956, where the roots of artificial intelligence can be traced, established certain goals: [2]

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.

When such goals could be achieved, the emergence of independent machine intelligence was assumed to be the logical next step. Most research in the sphere of artificial intelligence has an anthropomorphized approach; this is fair to say considering that some of the more popular work is being done in neural networks, the foundation and organizational structure of which emulates how neurons interact in the human brain. While the human brain is admittedly not optimized for intelligence, meaning the forms of intelligence exhibited by it are not necessarily the best and most efficient forms vetted evolutionarily, it is also true that there are several aspects of human intelligence that would do well to be implemented in intelligent agents. One such feature is metacognition.

Plato spoke of a system that allows for a learner, a teacher and an evaluator within our cognitive system, making this the first possible mention of a third aspect that evaluates our cognitive processes beyond simply learning and thinking. This was later defined as 'metacognition' or the "knowledge and cognition about cognitive phenomena" [3]. In current literature, this is generalized to 'thinking about thinking', allowing "one's own beliefs, reasonings, desires, intentions… [as well as] cognitive abilities, motivational dispositions, practical reasoning strategies" [4] to determine the extent and accuracy of one's thoughts and decisions.

Some definitions need to be briefly touched upon before establishing the theory. The 'dual process theory' (DPT) accounts for the processing of cognition through two different processes, referred to as System 1 and System 2. System 1 is the immediate response to a situation that "operates automatically and quickly, with little or no effort and sense of voluntary control" [5]. On the other hand, System 2 is the more deliberate system of thinking that "allocates attention to the effortful mental activities that demand it, including complex computations" [5]. The three significant distinctions between the two systems lies in:

1) the time between input and output,
2) the number of neurons activated in human cognition (here, we do not give in to chauvinism and instead allow for the generalization of 'neurons' to the smallest unit of cognitive ability, whatever that may be, depending on the agent of cognition. In the context of philosophy of mind, chauvinism refers to certain mental states being linked with physical elements limited to agents in which the state is confirmed [6]), and
3) the amount of energy required i.e., the entropy of the processes.

These three distinctions become crucial in the forthcoming argument of the need and implementation of metacognition.

It is essential to note here that this paper does not deal with the debate between DPT and single-process theory, and that it is only the features that are of import to the theory. Single process theorists believe that the difference arises due to degrees of cognition and that they are not separate kinds of processes altogether. Even so, both sides acknowledge the difference in terms of the three factors established above. Since resolving the debate "will not inform our theory development about the critical processing system underlying human thinking" [7], implying that beyond observing the outcome, there is nothing further to be understood about the architecture of cognition from establishing one over the other, this paper will instead proceed with a functionalist perspective in its use of DPT terminology. In the context of philosophy of mind, the concept of functionalism follows that "what makes something a mental state of a particular type does not depend on its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part" [8]. In this case, by taking a functionalist approach to System 1, we focus solely on the properties of such a system – properties which are observations of both processing theories regardless of their mechanisms – and therefore can be referred to by either theory with the same effect.

III. Proposal

The main focus of the approach in this paper is to significantly improve computational efficiency through two key features set up and handled by the metacognitive system:

1) Accessing a Database of Previous Computations:
   - *Creates and utilizes a centralized repository of past computations and problem-solving processes* - When encountering a new prompt, the system can search this database to find similar problems or computational patterns. This allows for faster resolution by reusing solutions, reducing the need for extensive recalculations.

- *Identifies the boundaries of expertise* - This is a critical extension of the strategy. In cases where a new prompt falls outside the range of knowledge or available solutions, the system can flag it and either redirect it to a more suitable computational resource or escalate it to human intervention. This ensures that the system remains efficient by focusing on problems within its scope while also handling edge cases appropriately.
- *Enables adaptive learning* - New computations can be refined or improved upon based on the knowledge and data collected from previous tasks, further optimizing the system's performance over time. This approach can also involve the redirection of prompts to relevant components or their respective field-specialized units, ensuring effective use of computational resources.

2) Segmenting Computation:
   - *Focuses on utilizing prior computations that share similar structures, properties, or nature* - By recognizing patterns and similarities between past and current tasks, computations can be reused instead of being repeated from scratch.
   - *Allows for parallel processing* - By breaking down tasks into specialized components, multiple processes can run concurrently, improving speed and efficiency without overwhelming any single, or the overall, available computational resource.

At the very outset, we must consider the kind of machine this theory can be implemented best through. Metacognition requires strong analysis tools that identify patterns, something best modelled by deep learning models. Schaeffer [1] integrates metacognitive processes into reinforcement learning frameworks. His model demonstrates that metacognition can be algorithmically implemented, enabling machines to detect their own suboptimal actions without external input. This provides a concrete example of how metacognitive functions can be translated into computational algorithms. Similarly, the proposed metacognitive system can be implemented via reinforcement learning, or other such machine learning methods, and can still be envisioned as independent of any computational cognitive system it is implemented in tandem with. The theory this paper proposes, the preprocessing metacognitive system (PMS), is the integration of the principles of metacognition and System 1 processes in computation, such that it may be able to provide a quick, immediate response before either proceeding in a certain manner or terminating computation altogether.

System 1, as the immediate response to an input, is heavily dependent on intuition in humans and "valid intuitions develop when experts have learned to recognize familiar elements in a new situation and to act in a manner that is appropriate to it" [5]. As such, intuition – and by extension, System 1 processes – can be reduced to an outcome of analysis, categorization and recognition of patterns which then manifests as a quicker response to situations that the untrained eye/mind would not notice patterns or decision-prompting cues in. More interestingly, however, is that this is only the case where one has expertise in the field, and that "when the question is difficult and a skilled solution is not available. . . [one] often answers an easier one instead, usually without noticing the substitution" [5]. This is a very human quality, a broad class of System 1 responses called 'heuristics' that are shortcuts established to answer questions with speed, regardless of accuracy. As mentioned before, speed-over-accuracy preference is also a persisting feature of machine cognition today.

If System 1-like processes were plainly implemented in machines, it would lead to false positives or false negatives, something that deep learning models are not exempt from. Like how people answer a particularly difficult situation almost immediately by looking at other cues that may not be relevant to the situation at hand, citing their intuition, machines also identify and call upon such misleading shortcuts when they are trained to observe patterns. This is most apparent when they mislabel or make errors in categorization based on certain other visible cues. Such models are widely implemented in deep learning models looking to identify images i.e., visual inputs and even in those, it is quite a challenge to train deep learning models to overcome these false results, requiring the dedication of many resources and an extensive database. To overcome these issues, the PMS becomes necessary. As previously mentioned, metacognition analyzes computational ability when it looks for patterns in reasoning to consider the matter of accuracy. Not only will it make it capable of terminating computation altogether if it falls outside the scope of such ability, but it will also carefully consider hindering features. Integration with System 1 principles of heuristics which are based on reasoning and skill-based cues allows it to create immediate output without resorting to actual computation, as well as avert the use of computation in cases that fall outside the ability of the machine altogether. This is what leads to trustworthy and robust outputs that only resolve things within the scope of the machine.

Thus, by acting as a system that makes use of analyzing tools before computation begins, the PMS can not only analyze the extent of, and patterns in computational ability to implement the benefits of a System 1 process – that of time, resource and energy efficiency – but also be able to determine the accuracy of the computation.

When looking into mapping the two systems onto each other, their inputs and outputs in the process become noteworthy. Metacognition acts as a separate system from cognition. The relation of DPT with metacognition can be understood such that the "default reasoning [System 1] is reasoning that precedes metacognitive control and intervening reasoning [System 2] is reasoning that follows metacognitive control" [9]. Let us consider the matter of what System 1 accepts as inputs and how its output correlates to the intermediary metacognitive system. System 1 accepts a task – be that identification, recognition, computation – objectively. On the other hand, metacognition processes 'reasoning cues' produced by System 1 while it develops judgements about the task.

At the outset, it is necessary to note that mapping one onto the other does not call for replacing one with the other

but instead a strategic mixture of their features, as has been necessitated above. If the metacognitive process were to be mapped onto a System 1 process, there would be a restructuring of I/O, and processing. It would develop such that the I/O of the PMS would be of metacognitive nature i.e., reasoning cues give directions and judgements, while the process itself follows the System 1 architecture, making it a matter of objective analysis of the cues and making connections using established heuristics to develop the directions as outputs. To further establish this idea, we consider the fact that the main computation, which will be of System 2-like nature, with deliberate control and use of more emphasized resources, can accept directional input from a boolean perspective, and have its own input passed on to it. Thus, to generate robust outcomes, the system would require a restructuring of the procedural hierarchy and not a replacement. That said, metacognitive systems would require their reasoning cues to come from somewhere, and this would come directly from an increasing database of the computation itself. The cycle would proceed as such: the metacognitive system is an 'onlooker' observing how computation works in a deliberate System 2 fashion. In the initial stages, it will be completely passive, only building a foundation of what kinds of problems the computation system faces, what kind of inputs generate a certain kind of input and what this can say about the processing ability of the machine. These will become the cues it calls upon when faced with new inputs, gradually identifying patterns in them and dealing with it accordingly.

Currently, most machine intelligence research targets actual computation, by accepting the input as is and working on it. This is a reasonable way to proceed when developing intelligence, but the implementation of this integrated system would allow for enhanced and accurate outputs by filtering out most inputs that have either already been computed before or fall outside the scope of ability, thus only utilizing the System 2-like deliberate computation to deal with new, solvable problems. It is also important to note that this system would not possess any cognitive abilities of its own but is entirely a pattern-recognition system. It identifies the requirements of the input and either matches it to a certain heuristic that has been established over repeated computations or gives a negative output altogether. In both cases, the system requires access to observe the computation itself of the machine it is working with – the data it works with comes from the machine itself and not from any external source of data. This is a similar process to how humans develop metacognition to form the basis of judgements, beliefs, reasoning, etc.

IV. Discussion | Evalution

Certain rebuttals may be raised against this theory; some of these will hereby be addressed. Perhaps the foremost one is the matter of metacognition not being a necessary component of human cognition, questioning its necessity in machine cognition. While it may be true that humans do not always rely upon metacognitive systems, the fact remains that it is full of erroneous judgements that develop from an oft flawed System 1 response. This arises due to the lack of logical integrity in the formation of heuristics. The principles of metacognition make it possible for the additional component of accuracy that is derived specifically from patterns analyzing computational ability.

The necessity of metacognition in current computation may still be doubted because there exist models today that function smoothly even without it, but it must be considered that none of these models have truly managed complex intelligence, and that some of the best work remains language models and neural networks. To reach a stage of cognition several orders of magnitude higher than the present, at the level of artificial general intelligence, there are certain other peripheral features that become necessary to ensure efficiency and accuracy, one of which is metacognition. Its necessity is based solely on the fact that intelligent machines require careful handling of resources, a 'wisdom' that develops through long-term analysis. The idea is like that of hierarchical, version-based intelligence, and could perhaps do away with the need of several versions of AGI if it can become a separate, constantly learning system entirely responsible for efficiency while the computational power increases separately. In simpler terms, wisdom deepens while intelligence improves.

There exist some common issues with models that are based on repeated learning, such as overfitting, where they provide excellent results with training data but fail with test data, because of overly adhering to certain features in the training sets and incorporating elements that would otherwise be deemed noise. While systems like regularization do exist to handle such issues, the prime issue with implementing such a system in PMS would be that it sacrifices accuracy for generalization ability. Instead, emphasis can be given to the idea of false positives and false negatives that were discussed previously. To expound upon the idea, the directions that will be given as output from the PMS will direct attention to the features of the input that hinder accurate computation, either due to lack of history in such spheres, or due to lack of clarity. In either case, the way it is handled by the computation once again becomes a learning for the PMS, allowing it to deal with such inputs differently and more effectively further on.

V. Conclusion and Future Work

To summarize, this paper explores how artificial intelligence can enhance its efficiency through the implementation of a preprocessing metacognitive system that acts as an automatic first-line-of-defense for all computation requests. It discusses how the proposed system can dynamically adapt its preprocessing strategies based on context, improving efficiency and accuracy in data handling. The system assesses its own preprocessing steps, identifies inefficiencies, and adjusts accordingly, mimicking human-like reflective thinking. This approach aims to enhance the adaptability, robustness, and self-improvement of further artificial intelligence systems in much more complex environments. The main aspects this paper furthers in terms of existing literature are:

- Self-Evaluation Mechanisms: Implementing algorithms that allow for assessing the quality of computations and decisions

- Redundancy Detection: Developing systems that reference a database of previous computations to identify and eliminate redundant operations
- Adaptive Processing Strategies: Creating frameworks that enable AI to adjust its computational strategies based on real-time self-assessment, optimizing resource allocation

This paper allows for there to be further objections to the conceptual explanation and details of the PMS; however, it also seeks to establish firmly the need for this theory. Without such a system, while progress will be made, it is less likely to be as immediately efficient and fast developing as it could be than if it were with it.

Further work includes developing and testing algorithms with the aforementioned features, drawing from the reinforcement learning models tested by Schaeffer [1], as well as expanding the range and examining other deep learning and neural network related systems in order to test for the most efficient implementation of the PMS. By integrating reinforcement learning insights and exploring diverse neural network architectures, future work aims to enhance the PMS's performance, ensuring its robustness and efficiency across a wider range of applications.

REFERENCES

[1] R. Schaeffer, "An algorithmic theory of metacognition in minds and machines", *ArXiv https://arxiv.org/abs/2111.03745*, 2021.

[2] J. McCarthy, M. L. Minsky, N. Rochester, and S. C. E., "A proposal for the dartmouth summer research project on artificial intelligence", *https://raysolomonoff.com/dartmouth/boxa/dart564props.pdf*, 1995.

[3] J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry", *American Psychologist*, 1979.

[4] J. Greco, "The social value of reflection", *In W. J. Silva-Filho & L. Tateo (Eds.), Thinking About Oneself: The Place and Value of Reflection in Philosophy and Psychology. Springer International Publishing*, 2019.

[5] D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[6] B. Gertler, "In defense of mind-body dualism", *In R. Shafer-Landau & J. Feinberg (Eds.), Reason and Responsibility.*, 2007.

[7] W. De Neys, "On dual- and single-process models of thinking", *Perspectives on Psychological Science*, 2021.

[8] J. Levin, "Functionalism", *In E. N. Zalta & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Summer 2023). Metaphysics Research Lab, Stanford University*, 2023.

[9] A. R. Dewey, "Metacognitive control in single- vs. dual-process theory", *Thinking and Reasoning*, 2023.