Toward Robotic Metacognition: Redefining Self-Awareness in an Era of Vision-Language Models

Daniel Leidner[1]

I. Introduction

Artificial Intelligence (AI) aims to replicate and even enhance human cognitive capabilities. However, AI researchers have primarily focused on planning and learning, leaving other concepts underexplored. One important concept is metacognition, which enables agents to understand and improve their own processes [1]. Metacognition is key to human self-improvement [2] and even the recovery of physiological limitations originating from stroke [3] or brain injury [4].

Recent advancements have challenged the traditional boundaries of achieving such human-level cognition through conventional methods. In particular, Large Language Models (LLMs) and Vision-Language Models (VLMs) have the potential to boost the development of cognitive architectures significantly. This blue-sky paper explores how VLMs can be leveraged to handle and avoid hardware failure and thus increase the resilience of future robots through metacognitive reasoning. With this we aim to showcases the transformative potential of metacognition as future research direction in robotics.

II. Related Work

Human cognition has always been considered a guiding principle for developments in robotics and AI [5], dating back to the robot *Shakey*, arguably the first cognition-enabled robot [6]. Cognitive architectures are by now a research field on their own [7]. An overview of cognitive architectures is found in [8]. One-third of the listed cognitive architectures support metacognitive features and thus self-awareness, yet only few are applied to robotics: The *Metacognitive, Integrated Dual-Cycle Architecture (MIDCA)*, focusing on task and motion planning issues [9]. The *Metacognitive Control Loop architecture (MCL)* enhances perturbation tolerance in reinforcement learning agents by integrating self-assessment [10]. Vinokurov *et al.* 's architecture evaluates action success using self-aware error monitoring [11].

Metacognition was recently identified as a key component toward generalized embodied intelligence [12]. However, until recently, implementing metacognition in robotics possessed a significant challenge due to one crucial obstacle: the extensive knowledge required to cultivate awareness about a robots own cognitive parameters and process, also known as metacognitive awareness [13], the first principle of metacognition, is challenging to attain through classical ontology-based knowledge representation and reasoning alone [14]. In contrast, this blue-sky paper proposes using VLMs to allow robots to independently access and utilize information about themselves, otherwise meant for humans only.

[1]German Aerospace Center (DLR), Institute of Robotics and Mechatronics, Münchener Str. 20, 82234 Weßling, Germany.daniel.leidner@dlr.de

III. Creating Metacognitive Awareness for Failure Comprehension through VLMs

LLMs and VLMs are not only able to access but also generate data that is otherwise difficult for conventional algorithms to process. This paper demonstrates this potential in alignment with the goals of the RECOVER.ME ERC Starting Grant project, which aims to replicate human metacognitive capabilities to increase resilience of future space exploration robots. In this regard it was just recently shown, that LLMs have the capability to generate Failure Mode and Effect Analysis (FMEA) documents for hazard analysis [15]. For this, a human analyst interacts with the LLM to support the explore possible hazard causes. In our study, we invert this paradigm to access similar information about potential hardware malfunctions of a Mars rover. The goal is to leverage the extracted knowledge for a structured analysis, which may eventually be conducted by a robot itself. We employ ChatGPT-4o as our VLM of choice[1] to explore how a robot may be able to access four common sources of information about itself:

Parse Schematic Images: We request the VLM to analyze an image of a general spacecraft subsystem fault tree [16]. The VLM successfully interprets the fault tree with a single prompt to "interpret this tree". Afterward the VLM can be queried to list potential failure sources by providing a description of symptoms. For example, a rapid energy depletion is correctly associated to fault of the solar array, the power distribution, or the battery.

Read Structured Tables: We provide the VLM with a FMEA table of possible failure modes for solar panels, as identified in [17]. The VLM is not only able to extract the listed failure modes from the table with a simple prompt to "interpret this table", it also makes reasonable adaptations to these modes considering conditions on Mars (i. e. by significantly reducing the likelihood of oxidation due to the absence of oxygen, while increasing the effects of wear and tear due to prevalent dust).

Analyze Research Papers: The information retrieval capability of ChatGPT-4o allows for advanced document comprehension. With a little help, the VLM is able to successfully retreive the percentage of wheel damage on the NASA Curiosity rover from the respective technical

[1]https://openai.com/index/hello-gpt-4o/

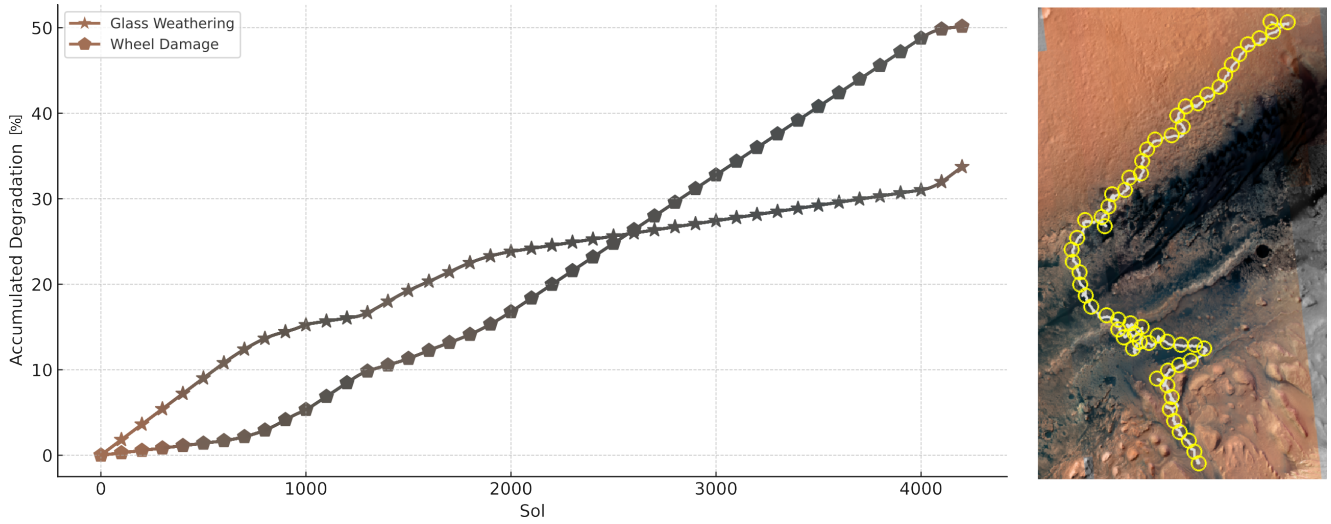

Fig. 1. Left: VLM-generated cumulative degradation of solar panel glass weathering from dust (stars) and wheel damage due to rough terrain (pentagons). Right: Map to extract terrain features (yellow circles) following the trajectory of the NASA Curiosity rover (white line) as of June 30, 2024 (sol 4210).

report [18]. By default, the VLM tries to extract the relevant information from the text corpus. However, as this information is not available in clear text, the VLM may hallucinate a number. As a fallback the VLM is queried to analyze visual cues from the figures of the report which provides more reasonable results.

Interpretation of Maps: The VLM is supplied with a map of the trajectory of NASA's Curiosity rover [19]. It is able to extract and analyzes color values along the 4210-sol trajectory. Upon request, the VLM is able to plot the prevailing failure modes of glass weathering due to dust and sand, as well as accelerated wheel damage as the rover traverses rough terrain considering the information retrieved from the previous prompts. The combination of facts is most difficult for the VLM, requiring manual refinement as well as additional prompts to show a legend, include markers, and better explain how dust and rocks may factor into the plot as seen in Fig. 1.

IV. Discussion

It is well known that LLMs and VLMs are prone to hallucinations and the observations made in this study are no different. It was the case that the VLM would sometimes resort to random values if it did not correctly interpret a prompt. For example, tracing the rover path on the map required multiple refinements to prevent the VLM from defaulting to a random route. Additionally, ignoring relevant information resulted in incorrect outputs. While Curiosity's average wheel damage is indeed about 50% as of sol 4210, *the wheel degradation plot is actually incorrect*. In reality, damage progressed much faster in the beginning of the mission as seen from the plot in Figure 29 of the technical report [18]. Although this circumstance is detailed in the text corpus, the VLM was biased to use the map information as specified in the prompt.

Consequently, we argue that it is of utmost importance to verify VLM-based hypotheses, especially in the context of failure handling and space exploration. In human metacognition, monitoring is a key concept to identify and avoid ill-advised decisions and update intrinsic parameters accordingly. This principle is referred to as metacognitive monitoring, which is a major part of metacognitive regulation [20]. Accordingly, interpreting the above-described resources solely by means of a VLM is not sufficient to achieve human-level metacognition. It is also necessary to be able to distinguish correct beliefs from incorrect assumptions.

The expression that *"there is no free lunch"* also applies to VLMs in the context of metacognitive awareness. The inability to differentiate between hallucinated and accurate responses necessitates verification through classical, model-based methodologies. This poses a significant challenge, as it is desirable to maintain the benefits of VLM-based knowledge extraction without redundantly replicating the same information using classical techniques. Metacognitive monitoring must be sufficiently abstract to avoid excessive overhead while still being detailed enough to detect potential errors. The RECOVER.ME ERC Starting Grant project will address this topic as one of its core research questions.

V. Conclusion and Outlook

The proposed approach was only tested through manually prompting ChatGPT-4o, yet the experiment demonstrates that VLMs have the potential to systematically analyze information that would otherwise be inaccessible to robots. However, while the latest VLMs interpret schemata, tables, research papers, and maps intriguingly well, classical methods remain essential to verify their output.

Nevertheless, we are confident that this innovation opens up a wide range of possibilities for future research. Robots will no longer be confined to curated data repositories; but access and interpret the same information available to humans. This includes not only the formal documents discussed in this paper, but also models, code, plans, and a plethora of other knowledge sources. Robots will soon achieve unparalleled metacognitive capabilities, and it is now up to the research community to seize this opportunity.

ACKNOWLEDGMENTS

Funded by the European Union (ERC, RECOVER.ME, 101116620). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. ChatGPT-4o was used to polish wording and generate figures where indicated.

REFERENCES

[1] J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry." *American psychologist*, vol. 34, no. 10, p. 906, 1979.

[2] N. Yeung and C. Summerfield, "Metacognition in human decision-making: confidence and error monitoring," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, no. 1594, pp. 1310–1321, 2012.

[3] J. Kersey, W. S. Evans, K. Mullen, A. Askren, R. Cavanaugh, S. E. Wallace, W. D. Hula, M. Walsh Dickey, L. Terhorst, and E. Skidmore, "Metacognitive strategy training is feasible for people with aphasia," *Occupation, Participation and Health*, vol. 41, no. 4, pp. 309–318, 2021.

[4] S. A. Makka, "Metacognitive skills training effect on cognitive function in traumatic brain injury patients: A systematic review," *Global Journal of Medical and Clinical Mini Reviews*, vol. 7, no. 2, pp. 085–099, 2020.

[5] A. Sloman, "20 varieties of metacognition in natural and artificial systems," *Metareasoning: Thinking about thinking*, p. 307, 2011.

[6] N. J. Nilsson, "Shakey the Robot," Tech. Rep., 1984.

[7] A. Lieto, M. Bhatt, A. Oltramari, and D. Vernon, "The role of cognitive architectures in general artificial intelligence," 2018.

[8] I. Kotseruba and J. K. Tsotsos, "40 years of cognitive architectures: core cognitive abilities and practical applications," *Artificial Intelligence Review*, vol. 53, no. 1, pp. 17–94, 2020.

[9] M. Cox, Z. Alavi, D. Dannenhauer, V. Eyorokon, H. Munoz-Avila, and D. Perlis, "Midca: A metacognitive, integrated dual-cycle architecture for self-regulated autonomy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[10] M. L. Anderson, T. Oates, W. Chong, and D. Perlis, "The metacognitive loop i: Enhancing reinforcement learning with metacognitive monitoring and control for improved perturbation tolerance," *Experimental and Theoretical Artificial Intelligence*, vol. 18, no. 3, pp. 387–411, 2006.

[11] Y. Vinokurov, C. Lebiere, A. Szabados, S. Herd, and R. O'Reilly, "Integrating top-down expectations with bottom-up perceptual processing in a hybrid neural-symbolic architecture," *Biologically Inspired Cognitive Architectures*, vol. 6, pp. 140–146, 2013.

[12] N. Roy, I. Posner, T. Barfoot, P. Beaudoin, Y. Bengio, J. Bohg, O. Brock, I. Depatie, D. Fox, D. Koditschek, T. Lozano-Perez, V. Mansinghka, C. Pal, B. Richards, D. Sadigh, S. Schaal, G. Sukhatme, D. Therien, M. Toussaint, and M. V. de Panne, "From machine learning to robotics: Challenges and opportunities for embodied intelligence," 2021.

[13] G. Schraw, "Promoting general metacognitive awareness," *Instructional science*, vol. 26, no. 1, pp. 113–125, 1998.

[14] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. K. Bozcuoğlu, and G. Bartels, "Know rob 2.0 - a 2nd generation knowledge processing framework for cognition-enabled robotic agents," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 512–519.

[15] S. Diemert and J. H. Weber, "Can large language models assist in hazard analysis?" in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2023, pp. 410–422.

[16] P. Z. Schultea and D. A. Spencerb, "State machine fault protection for autonomous proximity operations," 2017.

[17] P. Rajput, M. Malvoni, N. M. Kumar, O. Sastry, and G. Tiwari, "Risk priority number for understanding the severity of photovoltaic failure modes and their impacts on performance degradation," *Case Studies in Thermal Engineering*, vol. 16, p. 100563, 2019.

[18] A. Rankin, N. Patel, E. Graser, J.-K. F. Wang, and K. Rink, "Assessing mars curiosity rover wheel damage," in *2022 IEEE Aerospace Conference (AERO)*. IEEE, 2022, pp. 1–19.

[19] D. Blake, V. Tu, T. Bristow, E. Rampe, D. Vaniman, S. Chipera, P. Sarrazin, R. Morris, S. Morrison, A. Yen *et al.*, "The chemistry and mineralogy (chemin) x-ray diffractometer on the msl curiosity rover: A decade of mineralogy from gale crater, mars," *Minerals*, vol. 14, no. 6, p. 568, 2024.

[20] K. E. Lyons and P. D. Zelazo, "Monitoring, metacognition, and executive function: Elucidating the role of self-reflection in the development of self-regulation," *Advances in child development and behavior*, vol. 40, pp. 379–412, 2011.