Investigating Sense of Agency in Embodied AI Systems with Humanoid Robot NAO

Schiller, Adrian.

Studienprojekt
Humboldt-Universität zu Berlin

December 1, 2025

Abstract

This project investigates approaches to enhance the sense of agency in an embodied AI system controlling the mobile humanoid robot NAO, extending Yoshida et al.'s "mirror self-recognition test" originally designed for the stationary robot Alter3. The key challenge lies in NAO's full-body mobility, which introduces instability (falls) and spatial disorientation when vision is lost, hindering reliable self-assessment. We propose and evaluate three types of prototype extensions: Proprioception (real-time joint angle feedback in degrees), Direct motion execution (bypassing textual descriptions) and Developmental motion generation (phased unlocking of degrees of freedom). Experiments conducted in the Webots simulator (20 trials per setup, GPT-4o agent) demonstrate that proprioception significantly improves spatial awareness and reliability, yielding the clearest distinction between valid control and hallucinated agency. Direct execution in combination with proprioception reduces motion errors but increases false agency claims. The developmental approach combined with proprioception prevents early failures but requires proprioception for efficacy. Results indicate that modular systems with proprioceptive feedback are most promising for verifying agency in mobile robots, outperforming integrated or direct-execution paradigms. Future work should prioritise robust motion execution and enhanced image analysis.

1 Introduction

This project is based on the work of Yoshida et al. regarding embodied AI systems and self-perception by pairing state-of-the-art LLM GPT4o with the humanoid Alter3 [1, 2]. Following Gallagher et al.'s hypothesis that a sense of a self is based on a so called minimal self. [3]. The minimal self is defined as a pre-state or a building of full self-awareness which limits the perception of the self to the present moment instead of extending it temporally to form a narrative. The minimal self consists of a *sense of agency* and *sense of body-ownership*, meaning that the subject experiences that they have the ability to control their own actions (e.g. the movement of their body) and that their body belongs to them.

Previously, we developed a system to investigate sense of agency in the humanoid robot NAO based on Yoshida et al.'s work "Minimal Self in Humanoid Robot Alter3 Driven by Large Language Model" [2]. Yoshida et al. employed the ReAct-Agent architecture [4] to let GPT4 control the movement and head camera of a stationary robot named "Alter3". The robot was positioned in front of a mirror to answer the question of whether control over the robot can be verified. We replicated the system and employed it on the non-stationary robot NAO [5]. The main difference in these approaches is that Alter3 can mimic basic human emotions but only move the upper part of its body. In contrast, NAO can move its legs and hips but not mimic emotions. While our initial tests yielded some promising results, challenges became apparent. In this study project, we are investigating different approaches in extending the existing system to develop a *sense of Agency*.

In this paper, we will first discuss related work, our previous setup and challenges that came up in section 2. After that, we go over different approaches we used to extend the existing setup to overcome some of the challenges in section 3. We follow up with section 4, where we discuss our experimental setup and the results. Finally, we provide an outlook for future extensions and improvements as well as challenges in section 5 and provide a conclusion of this project in section 6.

2 Related work

In this section, we discuss the research that this project is based upon and go over our previous work and the challenges that came about.

2.1 Text to Motion

Since we employ language models to control the movement of a robot, one of the foundational concepts for this project is the generation of motion commands from natural language descriptions. Yoshida et al. demonstrated the ability of the LLMs to generate Python code to control the movements of robots in their 2023 paper "From Text to Motion: Grounding GPT-4 in a Humanoid Robot Alter3" [1]. To achieve this, they leveraged the power of few-shot learning alongside a detailed description of the joint axes of the robot Alter3 and instructions on how to format the output exactly. Alter3 is a humanoid robot that can move its upper body as well as its eyebrows and mouth, and can therefore mimic basic human emotions by facial expressions. The robot's axis joint positions are represented by identification numbers for each joint and values, which are the angles, normalised to a range from 0 to 255.

2.2 Minimal Self

Following up on their previous work regarding text-to-motion translation, Yoshida et al. implemented a system to experimentally test whether the LLM would be able to develop a so-called minimal self [2]. The concept of minimal self comes from behavioural psychology and was defined by Gallagher et al. [3] as a combination of the following two criteria. First of all, a sense of agency, meaning that one is in control of one's actions and realises that. Secondly, a sense of body ownership, meaning that one realises that the body belongs to oneself. To test these two criteria, they employ the "mirror self-recognition test" to investigate the sense of agency and the "rubber-hand illusion test" to investigate the sense of body ownership [2]. We will focus on the former, since our work focuses on the sense of agency. In the "mirror self-recognition test", the robot is placed in front of a mirror, and an AI agent is prompted to answer the question whether it has control over the robot's body. In the original setup, the agent can either perform one of three actions or provide a final answer if enough information is gathered to conclude the question. The three actions that can be executed are "capture image" (1) taking a picture with the robots camera, "image to text" (2) analyzing the captured image to to find out about the current position of the robot from the mirror's reflection or "generate motion" (3) executing a motion which is specified by the agent as a textual description and translated to motor code by the motion generation module. The experiments showed that the agent was reliable in judging whether control of the robot exists [2], resulting in an 80 per cent success rate in verifying control over the robot body.

2.3 Sense of Agency in NAO

In a previous project, we replicated Yoshida et al.'s "mirror self-recognition test" on the humanoid robot NAO. In contrast to Alter3, NAO cannot make facial expressions but can move its legs and feet. Our experiments showed limited success compared to the experiments of the original project. In two out of ten attempts, control over the entire body of NAO was verified, and in two out of the ten attempts, partial control was verified. The main problem in our setup was that the robot would fall over when attempting to move its legs and thus lose vision of the mirror, which is detrimental when trying to verify control over the robot in this setup. A major limitation in our setup came from the added degrees of freedom in NAO's leg movements compared to Alter3. The agent lacked spatial awareness when it lost vision of the mirror, and the motion generation module missed the context of the robot's current position, limiting its ability to execute movements accurately and recover from loss of vision. We also speculated that the translation of textual motion descriptions to actual motor code could lead to inconsistent movements and that the descriptions of the robot's position by the image analysis module could lead to information loss. Regarding the interpretation of the textual movement description, we also realised that the LLM would sometimes use values in degrees to describe the motion, which would be difficult for the motion generation module to execute since it relied on values in the range of 0 to 255 to specify the position of the axis joints.

3 Approaches

In this project, we intend to improve the existing setup with several prototype extensions to address some of the limitations that became apparent in our previous work.

Proprioception We implemented two types of proprioception to observe the current state of the robot's position, which is represented by the orientation of the robot's axis joints. The first addition to enable proprioception allows the agent to request feedback on the robot's joint orientations. For this, we implemented an additional module that outputs the state of the current orientation of the robot's joints in degrees and passes it back to the agent. We call this extension *Proprioception 1*.

Our second implementation of proprioception extends the motion generation module so that the robot's current axis joint orientations are appended to the AI prompt for the motion module. For this implementation, we decided to switch to a representation in degrees, as opposed to normalised values from 0 to 255, to be consistent with the values of the agent's proprioceptive feedback. We call this extension *Proprioception 2*.

Direct Motion Execution We further implemented an agent that is capable of directly generating the motor code for the robot in order to overcome translation issues with textual motion descriptions. For this approach, we provided the agent with all the necessary information about the robot's axis joints and the specifications regarding the formatting of the code. The motion generation module is not provided with the textual movement description, but with the actual code, which is directly sent to the robot to be executed.

Developmental Motion Generation Finally, we implemented a system which restricts the robot's movement at the beginning of each attempt and gradually allows for more and more degrees of freedom to allow the agent to develop an understanding of controlling the robot and prevent early failures due to falling over and losing vision of the mirror. To realise this, we separated the degrees of freedom for the robot's movement into four phases so that each phase unlocks additional movement options for the robot. Phase 1 only allows for the movement of the robot's head. After control of the head has been verified, the agent will go to phase 2, which allows for movements of the arms. Phase 3 allows for movements of the legs, and finally, phase 4 allows for movements of the entire body, unlocking all degrees of freedom. Only after verifying control over the currently allowed movements will the agent go to the next phase of motion verification. This extension naturally leads to attempts that take a lot more cycles of image capturing, analysis and motion generation, which increase the computation costs, since all previous steps are fed to the agent to provide the context for taking the next action. To counteract this problem, we decided to merge image capturing and image analysis into one module. This is justified by the fact that during our observations, the agent always performs an image analysis after capturing images.

4 Experimental Setup

We conducted the "mirror self-recognition test" with the standard setup, i.e., reproduction of the original project by Yoshida et al. [2], as a baseline and each of the implemented extensions. We also tested each extension in combination with the Proprioception 2. For each setup, we ran 20 trials to investigate the sense of agency (Group 1) and 20 trials in which we randomised the motor code to investigate whether the setups would lead to hallucinated agency (Group 2). All trials were conducted in the robot simulation environment Webots, in which a version of the NAO robot, as well as a mirror, is implemented by default. NAO was placed facing the mirror at a 75cm distance to ensure good visibility at the start of each run. All experiments were run using the state-of-the-art LLM GPT-4o to ensure consistency among the different setups.

4.1 Results

We split the final responses of the AI agents into three categories. The categories resemble the amount of body control recognised by the agents, so that the categories "Full", "Partial" and "None" mean that complete control, control over some body parts, and no control at all have been verified, respectively. The results were categorised by investigating the experiment's final outputs by hand.

The results of eight experiments are summarised in Figure 1. In our baseline trials, we observed 8 instances of fully verified control, 4 instances of partially verified control, and 8 instances where

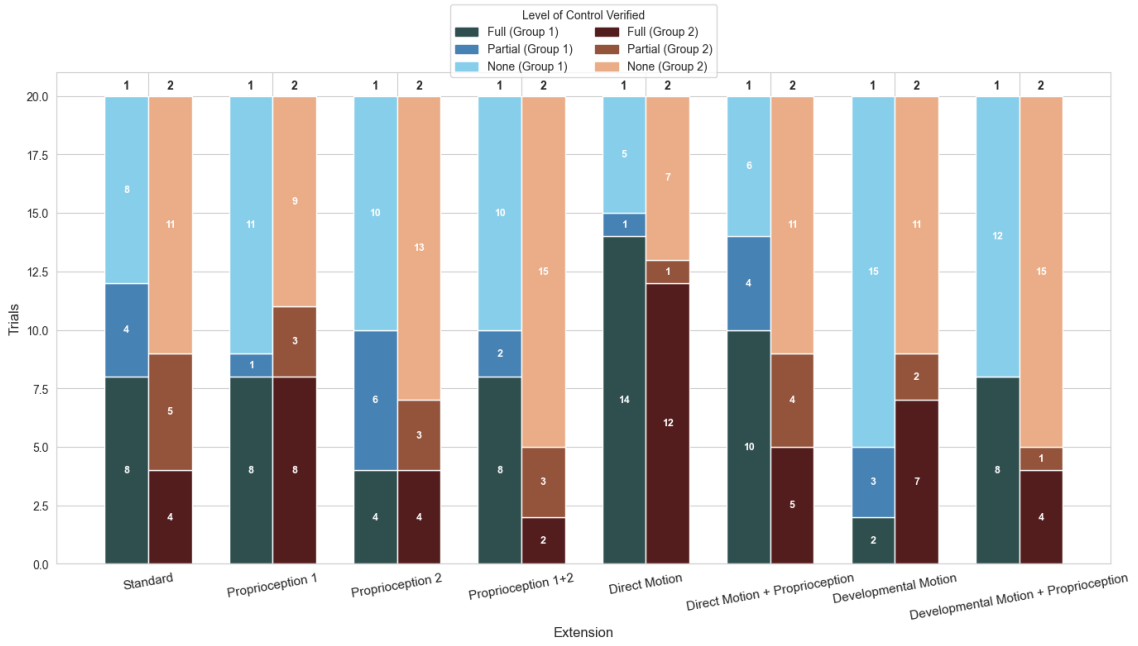

Figure 1: The experimental results are represented in the plot. The standard setup and each extension were tested in 10 trials. Group 1 represents the experimental trials of each setup, and Group 2 represents their control groups, i.e., the respective set-ups with random movements.

no control at all could be verified in group 1. We also observed that the robot fell over in 14 out of 20 trials. Out of the 6 trials in which the robot did not fall over, we observed 3 instances where the agent could not verify any control over the robot, 1 instance where partial control was verified and 2 instances where full control was verified. In group 2, we observed fully verified control 4 times, partially verified control 5 times and 11 times in total, where no control was verified. Upon reviewing the experiment logs, there was one instance where full control was verified due to poor logical reasoning by the agent, specifically because a series of movements was executed without verifying the mirror image for confirmation before claiming control. The proprioception extensions on their own did not bring improvements, as Proprioception 1 performed worse and Proprioception 2 showed similar performance to the baseline. Testing both proprioception-extension together, the number of trials in group 2 resulting in fully verified control decreased to 2 instances compared to the baseline, while that number stayed at 8 in group 2. Both group 1 and 2 included 2 fewer trials than the baseline, where partial control had been verified. Group 2 only contained 2 instances with fully verified control, for which one of the experiment logs revealed that only two movements, including one head and one arm movement, were executed, and the other one showed that it concluded to have full body control because of the proprioceptive feedback instead of the images.

In the setup where we let the agent directly generate the motor code, we observed the highest number of fully verified control with 14 instances and 1 instance where partial control was verified in group 1 but we also obserseved 12 instances of fully verified control and 1 instance of partially verified control in group 2, indicating that this setup tends to lead to halluzination in the agents reasoning. Combining this setup with the Proprioception 2 extension increased the discrepancies of group 1 and 2 by reducing the number of falsely verified controls in group 2 to 5, and keeping the number of correctly identified controls at 10. In both groups, the number of instances of partially verified control increased to 4.

Finally, the extension with developing degrees of freedom only led to 2 trials in which full control was verified in group 1 and 7 instances where full control was verified in group 2, though in combination with Proprioception 2, full control was verified 8 times in group 1 and 4 times in group 2.

4.2 Analysis

Analysing our results, we can see that compared to the baseline, we had the biggest improvements in the agent's ability to correctly determine whether it had control when the system was equipped with proprioception. The direct motion approach had very similar numbers of instances of fully

and partially verified control, suggesting that it is not a reliable method, though, in combination with proprioception capabilities, the system became more accurate. The developmental motion tool seems to increase the difficulty of concluding that control over the robot can be verified, since it ensures that, effectively, each body part has to be moved.

4.3 Discussion

By comparing the different setups with each other, we can see which directions we need to explore more to build a system that is fully capable of developing a sense of agency. First of all, the developmental motion extension, the direct motion extension, and Proprioception 1 did not show promising results on their own and would only work properly in combination with Proprioception 2. Our developmental motion and direct motion extensions showed major improvements when combined with a motion generation module that is aware of the current state of the robot's of the robot's position. Specifically, the developmental motion extension seems to prevent the agent from skipping to verify control over body parts and coming to a conclusion preemptively. Overall, the most promising extension was equipping the system with proprioception, which is not only indicated by the fact that it showed the biggest difference between group 1 and group 2, but also by the fact that we needed to include proprioception in the motion generation module for the other setups. These results indicate that we need to further focus on the improvement of spatial awareness and the motion generation module, suggesting that separating different functionalities into subsystems makes it more likely to develop a sense of agency as opposed to fully integrated systems.

5 Outlook

For further development of this project, we suggest looking into employing more sophisticated subsystems. Especially, the accurate execution of specified movements seems to be one of the crucial roadblocks to overcome. We suggest looking into research, such as Xu et al.'s 2025 paper "Realizing Text-Driven Motion Generation on NAO Robot: A Reinforcement Learning-Optimized Control Pipeline", where they successfully train a neural net to translate text to motion on a NAO robot [6]. Additionally, it should be looked into how image processing impacts the system, since we noticed that on a few instances during testing, the image analysis had the sides confused, e.g. claiming that the left arm was raised when the right arm was raised. Since the image analysis module does not have information about previous images and behaves independently from the rest of the system, it seems to fail in producing reliable and useful information for the agent.

Further, we want to briefly address more general approaches that we could not realise during this project. First of all, conducting the experiments with different LLMs (e.g. Google's Gemini, Anthropic's Claude) or completely local setups (e.g. Llama, Llava, Deepseek) used up too many resources, leading to long computing times (e.g. four to five minutes for one prediction) and unreliable outcomes (e.g., when it came to image analysis). We also recommend looking into multi-agent systems to establish more nuanced motion generation and image analysis that incorporates some sort of self-image that could be broadcast to all modules. This aligns with the idea of incorporating proprioception as a trivial way to maintain a body schema. Finally, we want to briefly address that this field of research is relatively new, so there are no standardised setups to assess the level of claimed body control or whether movement of any kind can be seen as control. Thus, further research includes more accurate definitions and reassessments.

6 Conclusion

This study project investigated different approaches to enhance the sense of agency in embodied AI systems, building upon Yoshida et al.'s framework of the "mirror self-recognition test" [2]. The main challenges came from the full-body mobility of the test robot NAO, including the possibility of the robot falling over and losing vision of the mirror and from spatial disorientation when vision over the mirror was lost. This project provides a foundation and clear direction for developing embodied AI systems capable of more sophisticated self-awareness.

Reproducablity

Our code is available at `https://github.com/afkadrian/minimal-self-study/`. A step-by-step guide to run the code is provided in the README.md file.

References

[1] Takahide Yoshida, Atsushi Masumori, and Takashi Ikegami. From text to motion: Grounding gpt-4 in a humanoid robot" alter3". *arXiv preprint arXiv:2312.06571*, 2023.

[2] Takahide Yoshida, Suzune Baba, Atsushi Masumori, and Takashi Ikegami. Minimal self in humanoid robot "alter3" driven by large language model. In *ALIFE 2024: Proceedings of the 2024 Artificial Life Conference*. MIT Press, 2024.

[3] Shaun Gallagher. Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences*, 4(1):14–21, 2000.

[4] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[5] KISUNG Seo and ALDEBARAN Robotics. Using nao: introduction to interactive humanoid robots. *AldeBaran Robotics*, 2013.

[6] Zihan Xu, Mengxian Hu, Kaiyan Xiao, Qin Fang, Chengju Liu, and Qijun Chen. Realizing text-driven motion generation on nao robot: A reinforcement learning-optimized control pipeline. *arXiv preprint arXiv:2506.05117*, 2025.