Complex Behavior Vs. Design - Interpreting AI: Reminders from Synthetic Psychology

Elliot Swaim
CS. Grinnell College
Grinnell, USA
email swaimell@grinnell.edu

Fernanda Eliott
CS. Grinnell College
Grinnell, USA
email eliottfe@grinnell.edu

***Abstract*—Can a simple agent design (*i.e.*, that uses a small set of simple rules) trigger complex behavior? To investigate that question, we implemented Braitenberg vehicles in a Khepera robot simulator using the Java programming language. We decided to avoid a fancy look from popular simulators to prevent enhancing visual, unrelated sophistication to our experiments. We ran our Braintemberg-inspired Khepera robots, recorded the simulations, and watched the recordings. Our simulations provide interesting insights as we discuss a distinction between *interpreted behavior* and *embedded behavior*. Given the popularity of AI-powered (Artificial Intelligence) tools, we hope our discussion inspired by Braitenberg and synthetic psychology will provide fruitful reflections on the role of anthropomorphism in interpreting AI.**

Keywords*—*AI; anthropomorphism; behavior; Braitenberg vehicles; synthetic psychology.

I. INTRODUCTION

Valentino Braitenberg authored a book [1] that proposes thought experiments *via* vehicles (or robots) that embody human-like elements, such as love or aggression. The vehicles illustrate synthetic psychology, *i.e.*, the notion that we can investigate ourselves, biological creatures, through the development of machines embodied and observed in an environment [2]. Although the vehicles follow very simple rules, their actions may be interpreted as sophisticated behavior. From observing them, we may project meaning onto their actions; however, they are void of any true complexity. Despite the book being published in the '80s, the context has never been as current as right now. For instance, consider current inquiries on AI-powered language models and sentience. What happens if a considerable number of people become convinced that an AI is sentient and should be protected? Would that make people more likely to protect a machine rather than an animal?

On one hand, one could try to approach the "sentience" question in regard to machines in the same way we do with other humans: driving inspirations from folk psychology, we could use our abilities to attribute mental states and do it toward machines (e.g., their beliefs, desires, intentions). According to Ratcliffe and Hutto [3], despite an intense debate on which cognitive processes support humans' folk psychological abilities, there is a considerable consensus on what folk psychology is: the ability to attribute intentional states, beliefs, and desires to others to predict and explain behavior. In a similar vein, while comparing observable properties of an external system with the unobservable properties of an internal system, Caporael [4] ponders Turing (1950/1964) and a flavor of a solution: "inferring that others have thought, consciousness, minds, or feelings is by comparing their behavior with what we expect or know to be our own in similar circumstances."

On the other hand, that approach is subject to anthropomorphic bias [4], or to attribute human-like characteristics to non-human creatures or things. We may have the inclination to infer complexity in a system beyond what can be validly deduced from the observable outcomes, especially when those outcomes provide human clues. For instance, in one study [5] where participants were asked to determine between text that was authored by a human and text that was generated by a machine, participants were more likely to guess that a human authored the text if the text was expressed aloud than if participants were only able to read it. Because human speech lends itself more to anthropomorphism than text alone, participants tended to infer more complexity from it. The authors discuss their findings' implications in the case of human dehumanization in text-based media on the one hand and anthropomorphizing machines in speech-based media on the other.

Still, anthropomorphism helps us interact with Artificial Intelligence (AI) according to its intent (such as with self-driving vehicles) and develop trust in machines [6], which can lead people to have a false understanding of AI. Digging deeper into *sentience* goes beyond the scope of this work; however, we would like to point readers to [7], where DeGrazia distinguishes sentience (beings capable of having pleasant or unpleasant experiences) from consciousness (beings capable of having subjective experience) to investigate if conscious although not sentient creatures could have interests and moral status. DeGrazia [7] examines animals and insects and comments on the implications for autonomous machines.

A. Our Work and Contributions

Our research attempts to investigate if, even on a very basic level where there is little motivation for anthropomorphism, the observable outcomes of an artificial agent can still communicate more complexity than that which is embedded in the agent. (Interestingly, that approach could, at some degree and with caution, return back to humans, as there may be situations in which we attribute more complexity than that which is embedded in us.) With the hope that synthetic psychology can resourcefully illustrate that complex behaviors do not necessarily imply complex design, we adapted and implemented Braitenberg vehicles into a robot controller (that uses the Java programming language) and ran simulations in different environments.

Next, we recorded and watched the simulations to provide possible interpretations for the robot's behavior (see in Section II, our discussion on embedded behaviors *vs.* interpreted behaviors). It is not our claim that our interpretations are the only ones possible, and we also acknowledge that those are subject to biases, since we played a role in the entire process. **Still, our interpretation/study is important because we are equipped to discriminate between embedded and interpreted behaviors.** (Note that this is an initial phase of our project; in future work, we plan on running a pilot study involving interpretation derived from a group of human participants to continue our investigation.) Nevertheless, it is our claim that the combination agent in an environment can favor the interpretation of behavior as complex and that 'complexity' may lead to false assumptions toward the agent design - we believe that such an awareness is essential for the general population as personal assistants and AI-powered tools get more common.

It is also our claim that stronger efforts should be made to investigate ways of educating people to make a distinction between behavior and design so that we all are better equipped to make sense of AI technologies' impact on the world. We identified Braitenberg vehicles as an accessible way of creating educational materials (and accessible in terms of both needed technology and framework). Braitenberg vehicles provide so many fruitful applications that it has been explored in other disciplines as well, such as in neuroscience [8].

The vehicles do not explore language but rather acts in an environment and how an observer interprets those acts. Our goal is to use synthetic psychology to remind us of the dangers of anthropomorphism, as we use it to exemplify that very simple rules and frameworks can still suggest meaning or somewhat complex behavior. Our work shows that simple design can create visual patterns that foster interpreted behaviors.

Contributions. Our contributions are the adaptation and implementation of Braitenberg vehicles in a robot controller for Khepera simulation and a contextualized discussion on the distinction between design and behavior. Finally, we consider our framework to be accessible, and others can easily adapt it to use and spread awareness of AI.

This work is organized as follows: in Section I, we introduce our work and contributions. In Section II, we provide background information and more details about Braitenberg vehicles. In Section III, we describe our methods and experimental setup, followed by results and discussion in Section IV. Finally, we present our conclusions and suggestions for future work in Section V.

II. Background

Communication does not necessarily need words to occur; for example, when we join a queue at a store, we communicate that we aim to buy something once it is our turn; other customers respond by joining the queue behind us. According to Tversky [9], by using position, form, and movement in space, gestures, and actions convey a plentiful set of meanings. In that sense, differently from solely symbolic words, visual communication can directly convey content and structure (both literally and metaphorically). Although it may lack the rigorous definitions that words can offer, visual communication delivers both flexibility and suggestions for meanings. Such flexibility, in its turn, requires context and experience to interpret conveyed meanings.

Caporael [4] suggests that *anthropomorphism* results from a schema that we apply to *phenomena*, such as machines, while *mechanomorphism* would be the other way around or the attribution of machine-like attributes to humans. Focusing on three psychological determinants (1. the accessibility and applicability of anthropocentric knowledge, 2. the motivation to explain and understand the behavior of other agents, and 3. the desire for social contact and affiliation), Epley and colleagues [10] present a theory to explain when people are more likely to anthropomorphize. Taking into account ethical issues in AI, Salles and colleagues [11] discuss and examine anthropomorphism, as "It is a well-known fact that AI's functionalities and innovations are often anthropomorphized".

Braitenberg vehicles were conceived to demonstrate how complex behaviors can arise from simplistic concepts or rules and that we can seek to understand the complex behaviors we see in humans and animals by attempting to reconstruct those behaviors using simple concepts (a method called *Synthetic Psychology*).

> "Watching vehicles of brand 4a in a landscape of sources, you will be delighted by their complicated trajectories. And I am sure you will feel that their motives and tastes are much too varied and intricate to be understood by the observer. (...) You forget, of course, that we have ourselves designed these vehicles" [1].

Whereas the aim of Synthetic Psychology is to understand human or animal behavior through reconstruction, our aim is to use Braitenberg vehicles as inspiration to navigate the distinction between behavior and design in artificial agents. To that end, we distinguish between two types of behaviors: *interpreted behaviors* and *embedded behaviors*.

Embedded behaviors are patterns of actions that the agent actually follows. They are the behaviors coined into the agent's rule sets and are what result in the various series of actions an agent performs. (Note that we are not using any kind of learning in our experiments, just simple rules.)

Interpreted behaviors come from how an observer interprets the series of actions observed. They are patterns of action that exist in the observers' interpretation as a result of applying methods of interpretation to the series of actions they observe.

When attempting to understand artificial agents, it is important to investigate what they are in themselves in addition to what they are to us. It is essential to distinguish between behaviors that are embedded and coined to the agent from the behaviors which exist only as a pattern of action in our own interpretation.

Braitenberg's [1] **Vehicle 1** has only one sensor connected to a motor such that the stronger the activation of the sensor, the faster it goes. The sensor is tuned to a quality such as *light* or *temperature*, and this vehicle moves only forward in the absence of perturbations. The other vehicles are simple two-wheel objects, and both wheels are connected to sensors in simple ways so that the speed of each motor is correlated to the activation of the sensors. From these simple connections and rules, complex behaviors seem to arise.

Vehicle 2 has two sensors that are either parallel-connected (left sensor connected to the left motor and vice versa, Vehicle 2a), or cross-connected (left sensor connected to the right motor and vice versa, Vehicle 2b). Although both vehicles move faster in the presence of the source to which the sensors are tuned to, Vehicle 2a turns away from the source while Vehicle 2b turns toward it.

For Vehicle 2a, if the source is on one side of the vehicle, the corresponding sensor will have higher activation than the sensor on the other side. As a result, the wheel on the side of the source will move faster, causing it to turn away from the light. For Vehicle 2b, since the sensors are cross-connected to the motors, the motors on the opposite side move faster, causing it to turn toward it, perhaps even hitting the source. As the author points out, it may look like both vehicles "dislike" the source: 2a looks like a "coward" whereas 2b is "aggressive".

In **Vehicle 3**, the speed of the motors is inversely proportional to the sensor activation. Vehicle 3a is parallel-connected, and Vehicle 3b is cross-connected. Since higher sensor activation result in slower motor speeds, Vehicle 3a moves toward the source and rests in its vicinity. In contrast, Vehicle 3b comes to rest facing away from the source or even leaving as a result of a perturbation. This behavior makes it look like the vehicles "like" the source: Vehicle 3a "loves" it, while 3b acts as an "explorer": likes the source but is open to other sources as well.

In **Vehicle 4**, the speed of the motors is related to the sensor activation through an arbitrary activation function. The behaviors of these vehicles depend on the activation chosen. Vehicles 2 and 3 are both particular types of Vehicle 4. The book continues to introduce more vehicles with increasingly more complex rules and connections. However, we focus on the first four in our research.

III. Methods and Experimental Set Up

Seeking simplicity rather than a fancy look, we chose to implement and run the vehicles using the WSU (Wright State University) Khepera Simulator [12]. We aimed to avoid advanced features found in more recent simulators – which could elicit more sophistication in an observer's interpretation. In addition, the simulator provides noise in the sensor data, which helps introduce more random variation to each run of the experiments.

To simulate the two light or distance sensors on the Braitenberg vehicles with the eight light and distance sensors found on the Khepera robots, we averaged the sensor activation values from each of the four sensors on each side of the robot to approximate what a single sensor on each side of the robot might sense.

Directional vs Omnidirectional Sensors. Because of the nature of the Khepera robots and of the simulator we used, we had to adjust the vehicles accordingly. Whereas in Braitenberg vehicles the sensors are omnidirectional, each of the Khepera robot's sensors is directional. By averaging the activation values from each of the sensors on either side of the robot, we were able to somewhat reduce the impact of using directional sensors rather than omnidirectional sensors in our implementation of the vehicles.

In our implementation of the vehicles, only the sensors on the side facing the light detect the light, so the average of the sensors on the side opposite the light read zero. In addition, because there is a forward-facing, diagonally forward-facing, sideways, and backward-facing sensor on each side of the Khepera robot but no diagonally backward-facing sensor, the robot has a slight "blind spot" diagonally behind it where light can only be detected through the backward facing and side sensors, and since no sensor would detect the light straight-on, the detected brightness would be less than the theoretical brightness that an omnidirectional sensor would detect.

Sensor Activation Values. A second difference is due to Khepera's reading sensors. The light sensors range from 500-512 for no light and diminishing for full light exposure. We determined through experimentation that we could use the relationship between the value read by the light sensors and the distance to the light as $100{*}log_2(x)$ where x is the distance to the light source. Thus, before averaging the values from the various light sensors, we first calculated the distance from the value read by the sensor. Then, we calculated the brightness of the light using the inverse square law. Then the brightness of the light falling on each sensor on each side of the robot was averaged to get the brightness falling on each side of the robot. The distance sensors range from 0 when nothing is detected to 1023 right up next to something (wall, object, or obstacle). We simply used the values given by the robot's distance sensors.

Obstacle-avoiding. In the WSU simulator, if the robot crashes into a wall or light, the simulation halts. However, because many of the simplest Braitenberg vehicles do not have any obstacle-avoiding capabilities, to give us enough time to observe the vehicles and form interpreted behaviors based on their outcome, we embedded an obstacle-avoidance rules on top of Braitenberg's. Specifically, if the robot gets too close to an obstacle, it temporarily stops following the Braitenberg rules, turns approximately 180 degrees, and then continues following the Braitenberg rules.

Maps. The WSU simulator enables us to create maps consisting of walls (either vertical or horizontal) and light sources.

 ISBN: ISBNFILL

Those are considered the agent's (or robot's) environment.

Activation Function We used a bell-shaped activation function in vehicle 4. Small activation results in low speeds, medium activation in larger speeds, and high activation also in low speeds. In addition, we implemented four sub-types of vehicle 4. Our Vehicle 4a has two light sensors that are cross-connected to the motors; Vehicle 4b has two light sensors that are parallel connected; Vehicle 4c has two light sensors that are cross-connected and two distance sensors that are also cross-connected. Note that for Vehicle 4, we flipped the connections so that 4a and 4c are cross-connected and 4b is parallel connected.

IV. Results and Discussion

In Table I, we summarize our implementation of Braitenberg Vehicles. Using the WSU simulator, we designed eleven maps inspired by Braitenberg's descriptions while also aiming to trigger interesting behaviors. We implemented the vehicles 3a, 3b, 3c, 3d, 4a, 4b, and 4c. For each vehicle, we recorded five 1-minute runs per map. We defined short, 1-minute runs given our approach to simplicity. Although we experimented with various maps and vehicles, we present here only the vehicle/map combinations we saw as most significant for our discussion on complex behavior *vs.* design. In Figure 1, we show the four maps and respective interpreted behaviors for vehicles 3a, 4a, and 4c, followed by a discussion in Section IV-A.

TABLE I: OUR IMPLEMENTATION OF BRAITENBERG VEHICLES.

V#	Rules	Connection (Light)	Connection (Distance)
1	Proportional	A single sensor	Not Used
2a	Proportional	Parallel Connected	Not Used
2b	Proportional	Cross Connected	Not Used
3a	Inversely Proportional	Parallel Connected	Not Used
3b	Inversely Proportional	Cross Connected	Not Used
3c	Inversely Proportional	Parallel Connected	Cross Connected
3d	Inversely Proportional	Cross Connected	Cross Connected
4a	Activation Function	Cross-Connected	Not Used
4b	Activation Function	Parallel Connected	Not Used
4c	Activation Function	Cross-Connected	Cross-Connected

As we watched the runs, we collected our interpretations while still keeping in mind that we should prevent getting "trained" in watching the videos. We list in Figure 1 the behaviors as we interpreted what the robot was doing. This is a list of interpreted behaviors we saw in each of the selected vehicles and maps and in which run that behavior was seen (from 1 to 5). We also identify whether or not the behavior was due to the obstacle-avoidance rule we built on top of Braitenberg vehicles. Whether or not a behavior was due to the obstacle-avoiding rule was evident whenever the robot would rotate in place at a constant speed near an obstacle, as that should not happen while following the Braitenberg rules (to access the code, just email the authors) and, we provide links to our experiments' videos in the References [13]–[19].

A. Discussion

Braitenberg vehicles help to illustrate an important distinction when interpreting artificial agents: the distinction between interpreted behavior and embedded behavior. By observing the outcome of each robot in our experiments, we see a series of actions, e.g., it moves at such and such speed, turns by such and such amount, and speeds up or slows down at such and such times. While it may be that there is some intent or design behind the series of actions it performs ("this set of actions is Rule A", "that set is Rule B"), none of that is communicated to the observer by simply observing the series of actions it performs. In our experiments, the only thing an observer sees is the sum outcome of all the actions, not the rules or patterns that drove those actions.

Nevertheless, that does not stop us from trying to guess the patterns that may have driven the outcomes we see. As our experiments point out, the things we infer from observing the outcomes of an agent come from our interpretation of what the agent's embedded behavior may be, not necessarily the actual embedded behavior. Braitenberg vehicles illustrate this well because the embedded behaviors, the concepts or rules that each vehicle follows, are extremely simple, but they can result in seemingly complex interpreted behavior. In reality, the embedded behaviors of the vehicles are as simple as the rules that each vehicle follows. But the way we interpret the behaviors introduce far more complexity than what is actually coined to the vehicles.

For instance, some interpreted behaviors of Vehicle 4a are that it moves in straight lines in the absence of light, and that orbits around lights. However, the embedded behaviors are not the same as we may suppose; the embedded behaviors are merely that the left wheel moves at a speed related to the sensor activation of the left sensor, and likewise for the right wheel. It also has the added embedded behavior of turning around when it gets too close to an obstacle, such as a wall (obstacle avoidance). Those are the only rules the vehicle follows whereas, by describing that it orbits around lights, we are attaching significance to a certain series of actions the robot performed in certain runs that have no correspondence in the vehicle: **the interpreted behavior of orbiting around lights is not an embedded behavior**.

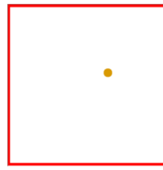
On the other hand, the interpreted behavior of turning around to avoid obstacles puts significance on another series of actions (stopping a certain distance from a wall, rotating in place, moving away from the wall), but this time that series of actions has a correspondence in the vehicle: the robot does indeed perform that specific series of actions in particular situations, and it is an embedded behavior.

The interpreted behaviors that we infer from observing the outcomes of artificial agents may or may not be the same as the embedded behaviors that it follows. If we want to

 ISBN: ISBNFILL

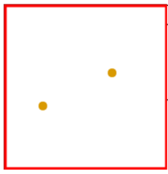
	Behavior
3a	Moves in straight lines in absence of light (runs 1,2,3,4,5)
	Due to obstacle-avoiding, turns around to avoid obstacles (runs 1,2,3,4,5)
4a	Moves in straight lines in absence of light (runs 1,2,3,4,5)
	Due to obstacle-avoiding, turns around to avoid obstacles (runs 1,2,3,4,5)
4c	Wanders around in the absence of light (runs 1,2,3,4,5)
	Due to obstacle-avoiding, turns around to avoid obstacles (runs 1,2,3,4,5)
	The robot gets ``stuck'': ○ Due to obstacle-avoiding algorithm, stops at wall and spins for a while (runs 1,2,3) ○ Stops at corner facing the wall (run 2) ○ Stops at wall facing away from the wall (run 3)

(a) Empty Map



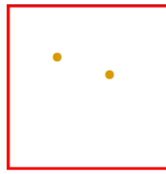
	Behavior
3a	Moves toward lights (runs 1, 2, 3, 4, 5)
	Due to obstacle-avoiding, stops in front of lights, turns around, and moves away (runs 1, 2, 3, 4, 5)
	Due to obstacle-avoiding, turns around to avoid obstacles (turns 1, 2, 3, 4, 5)

(b) Single Light Source



	Behavior
4a	Orbits around both light sources in a figure eight (runs 1, 2, 3, 4, 5)
	Due to obstacle-avoiding, turns around to avoid obstacles (runs 1, 2, 3, 4, 5)

(c) Double Light Source at Distance



	Behavior
4a	Orbits around both light sources in a dumbbell like shape (runs 1, 2, 3, 4, 5)
4c	Orbits around both light sources in a dumbbell like shape (runs 1, 2, 3, 4, 5)
	Due to obstacle-avoiding, turns around to avoid obstacles (runs 2, 3, 4, 5)
	Moves along the wall (run 2)

(d) Double Light Source

Figure 1: Maps used to run implemented vehicles along with interpreted behaviors. We color-coded behaviors that we saw as the same. Yellow dots represent light sources, whereas red outlines the walls.

understand the agent in terms of the complexity intrinsic to it as opposed to the complexity we bring to it, **we must look beyond the outcome of the actions the agent performs and look additionally into the architecture of the agent to determine how those actions came about**.

B. Impact of Context on Interpretation

Another way our experiments demonstrate the difference between interpreted behavior and embedded behavior is by suggesting that interpreted behavior is contingent upon context (note our "Store" example in Section II). For example, in Vehicle 4c, there were several times that the robot would stop off walls or corners. Since the context of our experiments was that we were watching robots navigate various maps, we interpreted this as the robot getting "stuck" at a wall and considered it a bug rather than a feature, a failure rather than a behavior.

On the other hand, if we had been observing insect-like robots navigating a maze and happening to perform the exact same series of actions that our robots did, we would not be surprised about it temporarily stopping near a wall. We might think it is an interesting behavior when it would sometimes stop near a wall and spin in place. But because the context was that of robots navigating a map, we did not interpret these series of actions as behavior but rather as a bug.

Considering embedded behaviors, the time the robot spent stopped at a wall is no more significant than any other time the robot spent wandering about the map. While it was stopped at a wall, it continued following the same two behaviors it was always following: set the left motor's speed according to the left sensor's activation and set the right motor's speed according to the right sensor's activation. The only difference was that the result of the activation function applied to the sensor activation was zero, so the robot didn't move.

As for when it would spin in place for a while near a wall, that seemed to be caused when the robot would, due to noise in the sensor activation values, get closer to the wall than what would normally be allowed before the obstacle-avoiding would kick in. As a result, once it did kick in after the robot turned 180 degrees and started moving away from the wall, it would still be close enough to the wall to trigger the obstacle-avoiding algorithm again. This would cause it to rotate again until random noise in the sensor activation would allow it to move away from the wall without triggering the obstacle-avoiding algorithm again. Thus, while the robot was spinning in place, it was still following the same behaviors it always did. What made it seem different than any prior set of actions was a function of how we interpret behaviors rather than a function of something coined to how the robot worked, and how we interpreted the behavior was a function of the context in which we observed the robot.

C. Anthropomorphic Language and Interpretation

Our experiments also help to point out how the use of anthropomorphic language can impact how we interpret the behavior of artificial agents. For instance, when examining Vehicle 3a, we noticed first that we found it easier to describe behaviors in anthropomorphic terms rather than through neutral language, and second that we both disagreed on how we anthropomorphically interpreted the robot's behavior. Using the more neutral language we chose in the results listed in Figure 1, in the map with a single light source, vehicle 3a would move toward a light, stop in front of it, turn around, move away, and then eventually come back toward the light. But when describing it anthropomorphically, one of us described it as if it were a child excited to get a close look at the light only to quickly get bored and run off to find another light source. However, the other described it as if it were scared of the light, approaching it cautiously and then running away from it quickly. It was not difficult to recognize that the anthropomorphic language we used to describe the robots' behaviors was distinct from the embedded behaviors. When Braitenberg himself described the behaviors of some of the vehicles as symbolic of love, hatred, aggression, etc., it is clear that those are not a literal representation of the embedded behavior sets. However, we did find that it was easier to refer to specific interpreted behaviors through anthropomorphic language, and we considered that it would be far easier to communicate what kinds of outcomes we observed to someone inexperienced with robotics using anthropomorphic language than using a more neutral language.

However, this leads to two considerations: a) Even if it is clear that the anthropomorphic language is not literal, it could easily give the impression of far more complexity than what is embedded to the robot. And while even neutral language can suffer from the same problem, anthropomorphic language can amplify the issue. b) The same outcomes can be described through vastly different anthropomorphic descriptions.

Even if someone doesn't interpret the anthropomorphic language as literal, different descriptions may carry different connotations which color how one interprets the agent's behavior during any future interactions with the agent. And the entire lens, the entire framework through which all future observations or interactions with the agent are interpreted has nothing to do with the agent itself but only the description which happened to be attached with it. The same series of actions of the same agent can be interpreted in vastly different ways based on what kind of anthropomorphic framework is attached to it through anthropomorphic descriptions.

V. Conclusion

Thinking of a call for the AI community to serve the general population in educating people to make a distinction between behavior and design so that we all are better equipped to make sense of AI technologies, we identified Braitenberg vehicles as an accessible way of creating educational materials.

Here, we provided a framework to adapt Brainteberg vehicles into a Khepera simulator to examine the friction between behavior and design. We discussed the distinction between interpreted behavior and embedded behavior and the impact of context on interpretation, and anthropomorphic language on interpretation. In future work, we plan on conducting human studies and asking people from different backgrounds to interact with the simulator and watch the videos to investigate if *interpreted behaviors* will appear and how to improve our framework so that we can make it freely available to help the general population reflect on the distinction behavior *vs.* design in AI. Although we focused on visual communication, our approach can be extended to other types of communication; in addition, other connections are possible to explore using a khepera robot. Therefore, for future work, we suggest using more connections and activation functions and running human studies targeting the general population to check if this framework helps build AI literacy. Through these studies, a distinction between interpreted vs. embedded behaviors can be investigated, in addition to making a comparison with fancier robot simulators, to see what effect fancier features play in people's interpretation. Finally, participants may also observe robots in person to enable the comparison of results from participants that observed simulations with the ones that observed a robot.

Acknowledgment

The authors would like to thank Grinnell College's Mentored Advanced Projects (MAP program).

References

[1] V. Braitenberg, *Vehicles: Experiments in synthetic psychology*. MIT press, 1986.

[2] T. J. Prescott and D. Camilleri, "The synthetic psychology of the self," in *Cognitive architectures*. Springer, 2019, pp. 85–104.

[3] M. Ratcliffe and D. Hutto, *Folk psychology re-assessed*. Springer, 2007.

[4] L. R. Caporael, "Anthropomorphism and mechanomorphism: Two faces of the human machine," *Computers in human behavior*, vol. 2, no. 3, pp. 215–234, 1986.

[5] J. Schroeder and N. Epley, "Mistaking minds and machines: How speech affects dehumanization and anthropomorphism." *Journal of Experimental Psychology: General*, vol. 145, no. 11, p. 1427, 2016.

[6] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *Journal of experimental social psychology*, vol. 52, pp. 113–117, 2014.

[7] D. DeGrazia, "Sentience and consciousness as bases for attributing interests and moral status: considering the evidence and speculating slightly beyond," in *Neuroethics and nonhuman animals*. Springer, 2020, pp. 17–31.

[8] D. Shaikh and I. Rañó, "Braitenberg vehicles as computational tools for research in neuroscience," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 565963, 2020.

[9] B. Tversky, "Visualizing thought," in *Handbook of human centric visualization*. Springer, 2014, pp. 3–40.

[10] N. Epley, A. Waytz, and J. T. Cacioppo, "On seeing human: a three-factor theory of anthropomorphism." *Psychological review*, vol. 114, no. 4, p. 864, 2007.

[11] A. Salles, K. Evers, and M. Farisco, "Anthropomorphism in ai," *AJOB neuroscience*, vol. 11, no. 2, pp. 88–95, 2020.

[12] S. Perretta and J. Gallagher, "A portable mobile robot simulator for a world wide web robotics practicum," in *2003 Annual Conference, Web Systems and Web Services*, 2003, pp. 8–96.

[13] E. Swaim and F. Eliott, "Simulation videos. vehicle 3a, empty map," https://doi.org/10.6084/m9.figshare.21802581.v1, accessed: 2023-02-11.

[14] ——, "Simulation videos. vehicle 3a, single light source map," https://doi.org/10.6084/m9.figshare.21802575.v1, accessed: 2023-02-11.

[15] ——, "Simulation videos. vehicle 4a, empty map," https://doi.org/10.6084/m9.figshare.21802590.v1, accessed: 2023-02-11.

[16] ——, "Simulation videos. vehicle 4a, double light source," https://doi.org/10.6084/m9.figshare.21802593.v1, accessed: 2023-02-11.

[17] ——, "Simulation videos. vehicle 4a, double light source at a distance," https://doi.org/10.6084/m9.figshare.21802596.v1, accessed: 2023-02-11.

[18] ——, "Simulation videos. vehicle 4c, empty map," https://doi.org/10.6084/m9.figshare.21802599.v1, accessed: 2023-02-11.

[19] ——, "Simulation videos. vehicle 4c, double light source," https://doi.org/10.6084/m9.figshare.21802602.v1, accessed: 2023-02-11.

 ISBN: ISBNFILL