Explainable Robot Design Based on a Robot Theory of Mind: A Web-Based Platform to systematically evaluate Robot Designs*

Kerstin Haring[1], Pilyoung Kim[2] and Daniel Pittman[3]

Abstract— **In this work, we investigate how the explanation of a robot's initially perceived capabilities is based on its surface-level clues and morphology. We explore how explainable robots tie into Robot Theory of Mind (RToM), a term we use to describe how people develop a mental model of a robot. We have developed a web-based platform to collect robot designs that are expected to correspond to mental states. We will train a set of Machine Learning (ML) models focused on feature extraction, validation of desired robot design attributes, and eventually use this as a tool to generate new robot designs targeting designs that provide an initial explanation about the robot's capabilities. We propose a series of neuroscientific studies to iteratively verify the outcomes from the data collection and the ML models training on data provided by the Build-A-Bot platform.**

I. Introduction

The explainability of robot systems depends on the ability of people to reliably predict the robot's abilities [1]. This means that people form certain expectations about a robot [2] and they form a theory of mind (ToM) of the robot [3], [4]. A theory of mind (ToM) is the cognitive capacity to attribute minds to others and describes the ability to perceive mental states in others [5]. ToM or mental state reasoning represents a critical cognitive input for behavior explanation, action prediction, and moral evaluation [6]. That would mean that they would ascribe mental states to a robot that go beyond the actual robot's abilities. This could manifest, for example, in the human experiencing empathy or emotions towards the robot, the attribution of beliefs, goals, and desires to the robot, and ascribing mental states like agency and experience [7] to the robot.

One research approach that immensely contributed to the Social Robotics community is the research on how robot behaviors lead humans to ascribe mental states. This workshop contribution, however, takes an even more foundational approach and focuses solely on the impact of robot morphology on forming of a Robot Theory of Mind (RToM). One of the first things humans turn to when forming a mental model of a robot are the surface-level cues they experience when first seeing a robot. Creating an unambiguous first impression could contribute to more effective communication with the robot, [8] and a clear mental model of the robot is elicited if the robots match their task [9]. When investigating RToM in the context of robot morphology, we want to be able to determine which morphology features of a given robot design are most important in whether a human forms a RToM, which features elicit a clear mental model of the robot, and also what specific features or the combination thereof communicate about a robot's abilities.

To achieve this, we developed a platform where users design a robot towards a given attribute. For example, a user is asked to design a robot that can act autonomously or experience joy. This reverses the current paradigm where participants are asked about an existing robot (design) and what their perceptions, expectations, and discerned capabilities of the robot are. It is expected that the resulting robot designs shed light on what features a robot design should have to display a certain robot capability.

II. Related Work

Theory of Mind is a social-cognitive skill that involves the ability to understand that other people's thoughts can be different from your own [10]. It is called a theory because we have no direct way of knowing exactly what another person might be thinking, so we rely on our theory that we develop based on their appearance, behaviors, and what we know about them. If we assume that people form a Theory of Mind of a robot, they have mental state concepts of the robot, such as "believe," "know," "want," and "see," and use them to predict and explain robot behaviors. A human with a Robot Theory of Mind believes that mental states play a causal role in generating behavior and infers the presence of mental states in robots by observing their appearance and behaviors [11].

If people develop a RToM they would also, at a minimum, develop empathy towards robots. Empathy refers to emotional awareness of others' feelings, or in case of robots, simulated expressions of feelings. While it is an emotional reaction that is appropriate given another person's mental state, it is just one component of ToM. ToM overall is a more complex cognitive ability of grasping the other person's perspective [12]. Prior research has assessed that people indeed show empathy towards robots [13], [14], [15], [16], [17] and that robot morphology impacts perceived empathy [18]. However, it is unclear which exact robot morphology features elicit human empathy.

Also, if humans were to have a RToM they also would anthropomorphize robots, i.e. interpret robots in terms of human characteristics and emotions. In essence, they would be

*This work was supported by the University of Denver's Professional Research Opportunities for Faculty (PROF) under grant # 142101-84994

[1]Kerstin S. Haring is with the Ritchie School of Computer Science and Engineering, University of Denver, Denver, CO 80210, USA, kerstin.haring@du.edu

[2]Pilyoung Kim is with the Department of Psychology, University of Denver, Denver, CO 80210, USA pilyoung.kim@du.edu

[3]Daniel Pittman is with the Department of Computer Sciences, Metropolitan State University of Denver, Denver, CO 80204, USA dpittma8@msudenver.edu

humanizing the robot. Research has come to the conclusion that people anthropomorphize robots, especially those that show a more human-like morphology [19], [20], [21], [22], [23], [24]. Research has also shown that robots are perceived to "see", "want", "know" [25], [26], and to be trusted [27], [28], [29]. Further evidence that humans might indeed form a RToM is that when the tests used to evaluate a Theory of Mind for humans (white lie test, behavioral intention task, facial affect inference, vocal affect inference, and false-belief test) are applied to robots, it has been shown that people implicitly assign mental states to robots [30].

It seems that robots might have all the necessary components that would lead humans to create a Robot Theory of Mind that is, at least initially, based on the robot's morphology. We hypothesize that we can determine which robot design features de- and increase a RToM and that we are able to predict what mental states are ascribed to a certain robot morphology.

III. Methodology

In order to prove our hypothesis, we need to explore a wide range of characteristics of robot design that entail both higher order cognition (e.g., rationality and logic) and emotion (e.g., feelings and experience [31], [32]. These two core capacities are mapped to two dimensions of the perception of the mind: agency and experience [7]. It has been demonstrated that people automatically evaluate a target's mind along these two dimensions [7], and non-human targets can be living entities such as animals [33], and non-living entities such as robots [32], [34].

Traditionally, investigating the effects of robot morphology involved survey-style research in which an existing design is presented to a user. The user is then asked to evaluate the design in terms of a certain mind perception attribute. Although this style of research works well for investigating individual designs, it is difficult to generalize the findings to other existing or new robot designs, due to the low number of designs evaluated as compared to the total number of existing and new designs. For example, the currently largest database of robot designs classified by their human-like appearances contains only about 250 existing robots [35]. This makes it difficult to evaluate the level of Robot ToM (RToM) that a new or existing unstudied design may elicit in a human.

The Build-A-Bot platform does not only intend to research the explanability of robot's perceived capabilities, it also aims at making predictions about new, not yet existing robot designs. With a large number of robot designs tied to an attribute, we can employ Machine Learning (ML) algorithms to evaluate the causal relationships of robot morphology and mind perception. In addition, currently developed databases and Machine Learning models are often not verified with independent methods for their validity. We propose a series of neuroscientific studies to iteratively verify the outcomes from the data collection on robot morphology and the Machine Learning models. Neuroscience technology such as functional near-infrared spectroscopy (fNIRS) has been used to understand neural responses to social robots as an implicit response evaluation [36], [37] and will be used to validate our approach while providing additional insights in novel ways of measuring interactions with social robots.

The Build-A-Bot platform addresses the challenge of significantly increasing the number of robot designs by a wider range of designers than currently is the case. We created a web-based research platform prototype where users can create any robot design they deem appropriate to a given prompt and use the Unity-based drag-and-drop interface to assemble a robot.

To build a model that learns from the user-based input of robot designs, we are looking to use the targeted attributes and the associated robot designs created on the Build-A-Bot platform to train a set of Machine Learning models focused on feature extraction, validation of desired RToM attributes in proposed robotic designs, and eventually as a tool to generate new robot designs targeting a given RToM attribute.

An additional challenge we identified is that we currently lack an independent assessment of Machine Learning models. We will use novel methods and technologies in neuroscience (i.e. fNIRS) to validate our ML models. This will serve as a novel metric to assess the validity of the data generated from the platform and the ML models.

IV. Current Results

A fully functional prototype of the Build-A-Bot design platform is available online at https://www.dubuildabot.com. We explored several options before selecting the Unity 3D game development platform [38] as the basis for Build-A-Bot. The Unity system allowed for rapid development and had the benefit that our platform could be deployed using WebGL [39]. Through several design iterations, we created a drag-and-drop system that allows users to combine 3D robot parts in any manner that they see fit (see Figure1. In order for the user to design for a specific attribute, they are prompted with *challenge card* (see Figure 2. The included *number of parts* requirements are needed as a measure of complexity for the machine learning models and the neuroscience portions of the project.

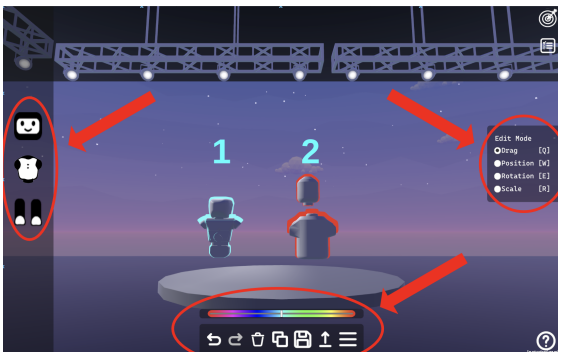

Fig. 1. A screenshot of the robot building tool with the part selection for drag-and-drop and the left side, a demonstration of an attached (1) and unattached part (2), the edit gestures on the right side, and the coloring and action menu on the bottom of the screen.

Complexity is defined as the number of variables that are related in a cognitive representation of a robotic design.

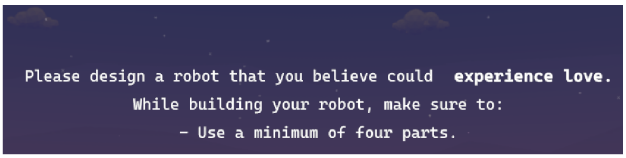

Fig. 2. An example of a *challenge cared* presented to a user before they start creating a new robot design.

Complexity has been shown to influence performance in ToM tasks and depend on the functioning of the frontal lobe, the area we aim to measure for RToM [40]. We obtain a complexity measure for a given design by assigning several attributes, including organizational information such as a category and subcategory label, as well as a complexity score for each new robot part created for the platform. The complexity score for individual parts allows us to calculate an overall complexity score for a robot design. We then use this score to help us group different robot designs by complexity.

To increase universal usability of the platform to a broad spectrum of users, we implemented several Human-Computer-Interaction (HCI) best practices [41]. For example, we created a tutorial that walks a first time user through the process of creating a simple robot design. This gives a user the opportunity to become more familiar with the platform before creating a robot design. We also used a drag-and-drop mechanism to select and position parts, manipulation via mouse or keyboard of the individuals parts (see Figure 3, and visual and sound effects when parts attached to each other (i.e., yellow sparks and a snap-like sounds effect). The interface also allows to adjust to a users environment with a light and dark mode. However, there are several areas in this prototype where we still can improve on universal usability. For example, the current *challenge cards* are text-based and we are in the process of evaluating several icons to support the understanding of the challenge cards for non-native speakers or illiterate users. We also are employing iterative user-testing and subsequent re-design of the website and the robot builder interface where we change the interface based on qualitative and quantitative user feedback.

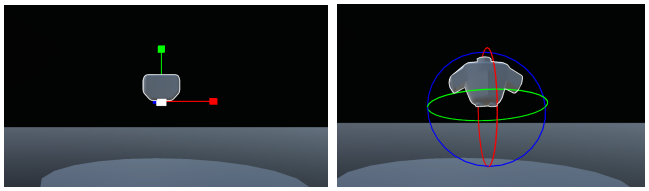

Fig. 3. Examples of scale and rotation handles for modifying existing parts.

When developing the Build-A-Bot platform, we realized the need for an effective way to manage the data collected from the platform. To achieve this, we are creating a researcher dashboard through our website that our team can use to perform queries against our database of submitted robots. This will allow us to timely look for patterns within different challenge cards presented to the user, such as one part being used frequently between different designs for a scenario. The dashboard will also allow us to incrementally add and remove challenge cards that will be presented to users to broaden data collection.

Since Build-A-Bot is hosted on a publicly accessible website, we can use a standard web server and database configuration to save data on robot designs. As the user builds a robot design on the platform, we incrementally save the changes they made. Each change, for example a change in the position, rotation, scale, or color of a part, is recorded in a JSON file. We are also tracking the order in which parts are they were chosen as is crucial to identify which components might be more important for users or if our interface introduced confounding variables by introducing an order of parts. For example, if we find that a certain part is frequently used in designs targeting a certain attribute, this would be an indicator that this part is important. If we find that a certain set of parts are always used regardless of the prompt given to the users, however, we need to reconsider the interface design to make sure users aren't being biased towards selecting those parts. The data is also used to build a tree-like structure of the path the user took to their final design, and we anticipate using these models to represent how different input variables can be used to predict a target robot design.

Once the user has indicated that their design is complete by submitting it, they are no longer able to edit it as we want a snapshot of the design that the user believed best represented the target attribute. A user however can copy their own existing designs and make edits to the copied version. Users can see all of their own robot designs in their account. After the user has completed their robot design, a screenshot of the robot design is taken and stored to a database. We provide the user with a 360-degree turntable animation of their design that they can download and keep. An example of this is shown in Figure 4. The robot design files themselves are stored as 3D objects and we plan to make them available as STL files to the user in future iterations for the platform to facilitate 3D printing of their robot designs. The screenshot provides a simple visual representation of the design that can be quickly re-visited during analysis or presented to users on our website. This screenshot can be used as an instrument for our neuroscience assessment as stimuli and in our Machine Learning models to facilitate learning via Convolutional Neural Networks (CNNs).

Fig. 4. An example of the 360-degree turntable animation for a created robot design

The other data we collect through the Build-A-Bot platform, such as the JSON data representing the robot and the path taken to create it, will also be instrumental in the other aims of our work. We plan to use this data to build ML models that allow us to create new designs targeting a

specified RToM attribute. ML models like these require a significant amount of data, which the Build-A-Bot platform is built to provide.

V. Discussion

We created a fully functional web-based platform to collect robot designs that are expected to correspond to mental states that play a causal role for RToM [11] and that offer an initial explanation of a robot's capabilities to the user based on robot surface clues and its morphology. To evaluate RToM based on the user input of robot designs, we need to have as much information as possible on the user's design choices. This will be achieved by tracking a variety of measurements for further analysis while the user is building a robot design via the platform (e.g., resulting designs as images, user actions as json file).

A. Limitations

The Build-A-Bot platform is currently limited by a comparatively small number of 3D robot parts available to choose from. While there are a number of pre-built robot designs available on the internet that we could use, very few of them are created in a way that they could be used as versatile part of the robot (i.e., a part could be a torso, hands, arms, legs, head, etc.). This would significantly limit the creativity of users and the models we can build. Since we are creating a platform where users can assemble parts on their own, a pre-built design represented as one model is of little use to us. We therefore are creating all the 3D models of robot parts on our own; however, this is a time-consuming process. We are continually working to increase the number and variety of parts available on the platform. Also, we are looking into enabling an advanced function where user-submitted parts could potentially be used. In either case, we still would need to validate the parts for appropriateness and polygon count, as parts with very high polygon counts have been found to cause the platform to become unresponsive.

Another limitation is the current lack of iterative design, user testing, and expert review. While we have run preliminary user testing as well as a first round of qualitative and quantitative user evaluation, we can only evaluate the universal usability of the platform after including a broader demographics of users (e.g., include children, adolescents, and elderly), and after including expert reviews of a (mostly) bug-free version of the platform. After the recent focus on rapid and iterative development of the platform, a good next step at this point seems to be to focus on improving the overall interface by hypothesis-driven testing. The enable universal usability is expected to correlate with the quality of the data that we will be collecting for the analysis of an RToM and how to develop explainable robots by design.

VI. Future Work

A. Improving Robotic Designs Using Human-Centered AI

Human-Centered Artificial Intelligence (AI) is a specialization focused on bridging the gap between humans and machines by developing intelligent systems that can understand how humans perceive and interact with the world around them. As part of this project, we will develop a Human-Centered AI approach to processing the design data collected as part of the Build-A-Bot platform. Specifically, we will be turning to Machine Learning (ML) to help us better understand how the perception of a robot mind and the perception of the explanation of robot capabilities are causally related to robot design. By creating machine learning models capable of predicting the RToM perception of a given design, we can create robots with designs targeting a given RToM attribute. In order to better understand what kinds of ML models are best suited for identifying robot design features tied to explainable robots, we will be experimenting with combinations of data preprocessing techniques and deep learning configurations to find ideal ML pipelines suited to explain RToM and how to develop robot designs that increase the explanability of robot capabilites.

Our data consist of both images of the created robot designs and low-level model information, including features such as what parts were selected, where it was attached to the design, and what rotation or scale was applied. We will create models based on the pixel information provided by the screenshots of the images as well as models based on the robot design's composition, and compare the accuracy of these models to see which can be used as a better predictor explanability for a given design. Our hypothesis during this comparison is that the pixel values can serve as a better predictor of whether a new design conforms to a given RToM target, while the design composition data will serve as a better training dataset for models used to generate new designs. We will also look at grouping models by the relative complexity of the training examples to test the hypothesis that ML models focused on a given complexity will be more accurate than ones trained with a mixture of complexity values.

In order for our human-centered AI system to continuously improve itself, all models developed as part of the project will be updated as new designs are submitted to the system. The models will be made available for public use via our project's website, allowing users to experiment with testing their own designs, generating new designs on the fly, and contributing to the research community by increasing available broad-spectrum robot designs.

A critical aspect of our project is to build trust with the community that the predictions and designs created by our ML models can be trusted. To help engender this trust, we will take the predictions and designs created by our system and validate them using neuroscientific experiments. By comparing the data collected through these experiments for a given design with a known baseline set of values, we can determine whether the response of a given user matches the predicted response of our ML models. This provides us with an objective measurement of precision that can be used to build trust in the results and creates a novel way of assessing the explanibility and perception of robot design.

B. Validating RToM Perception Using Neuroscience

Independent of the results of the platform design, we are developing implicit and explicit measures to build a model of the perception of the robot mind. Novel metrics are used to assess and verify the results of the platform robot designs, as well as to assess and refine the ML models. To date, robot mind perception research is in its infancy and has no models based on a large dataset and implicit measures to verify its validity. Additionally, no prior research attempted to use implicit measures to link robot design features with robot mind attribution. The interactive testbed we are developing includes explicit measures (e.g. questionnaires), as well as neuroscientific measures (i.e. fNIRS). Functional Near-Infrared Spectroscopy (fNIRS) explores functional activation of the human cerebral cortex through optical topography. It is noninvasive, silent, low cost, portable, allows participants movements, and has good temporal resolution, which is highly desired as we investigate responses to stimuli (see Figure 5). Studies using physiological measures have shown that EEG can pick up on differences in neural responses to pain stimuli for cross-racial empathy [42], that fMRI can show emotional and neural processing differences when observing human-human vs. human-robot interactions [43], and that Functional Near-Infrared Spectroscopy (fNIRS) is suitable to detect modulation of empathy [37]. fNIRS has also been mentioned to be a physiological method necessary in future HRI studies to determine how social robots should be designed to best perceive user needs [44]. In order to investigate how fNIRS can help reinforce our findings into RToM produced from the machine learning models training on data provided by the Build-A-Bot platform, we are planning to integrating prior work on fNIRS and HRI to build a new experimental approach in order to establish a proof of concept for measuring RToM. It is strongly expected that these insights on mind perception of a robot correlate with the way users explain a robot's expected capabilities and behaviors.

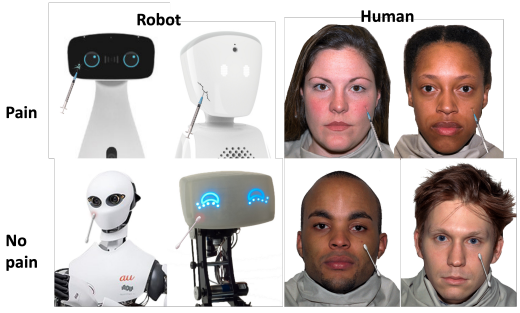

Fig. 5. Stimuli examples for a planned preliminary study comparing pain vs. no-pain and human vs. robot conditions.

Our study will focus on investigating an event-related reaction of participants' empathy reactions to human and robot faces displayed in a painful condition (i.e., with a needle penetrating the skin) and in a nonpainful condition (i.e., with a Q-tip touching the skin, see Figure 5). A repeated measure ANOVA will be performed to determine the effect of the touch condition on the fNIRS results, which will allow us to determine whether there is a difference in activation between a painful touch and a pleasant touch. This will help us determine if fNIRS can provide insights into RToM that we can use as a validation method for the findings from our machine learning model, and provide a novel experimental method into RToM that can lead to significant new insights.

VII. Conclusion

We have successfully developed a comprehensive experimental design that has the potential to significantly increase the knowledge of how people develop Robot Theory of Mind (RToM) and how to use this knowledge to design robots whose capabilities can be explained by the user based on their surface-level cues. We created a web-based platform that will collect a large amount of robot designs associated with a mental state. We will be able to determine what mental states are ascribed to robots and how a robot needs to be designed to display a certain mental state and trigger a certain expectation or user explanation. In the future, this work will utilize machine learning and neuroscience to significantly contribute to knowledge in each respective field and give insights on a more comprehensive assessment of interactions with social robots.

ACKNOWLEDGMENT

We are grateful to our students Itzel Bailon, Jenna Chin, Mike Blanding, Benjamin Dossett, Marley Bogran, Josh Ellis, Ryan Guyton, Yechan Han, Izzy Johnson, Weston Laity, Robel Mamo, Hector Armando Rodriguez, Ashley Sanchez, Jordan Sinclair, Marta Sinitsina, Maisey Toczek, Nicole Train, Kelly Trujillo, for their work on DU Build-A-Bot.

References

[1] S. Thellman and T. Ziemke, "The perceptual belief problem: Why explainability is a tough challenge in social robotics," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 3, pp. 1–15, 2021.

[2] K. S. Haring, K. Watanabe, M. Velonaki, C. C. Tossell, and V. Finomore, "FFAB-The form function attribution bias in human-robot interaction," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 843–851, 2018.

[3] A. M. Rosenthal-von der Pütten, N. C. Krämer, L. Hoffmann, S. Sobieraj, and S. C. Eimler, "An experimental study on emotional reactions towards a robot," *International Journal of Social Robotics*, vol. 5, no. 1, pp. 17–34, 2013.

[4] F. Hegel, S. Krach, T. Kircher, B. Wrede, and G. Sagerer, "Theory of mind (tom) on robots: A functional neuroimaging study," in *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 335–342, IEEE, 2008.

[5] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind"?," *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.

[6] L. Young and A. Waytz, "Mind attribution is for morality," *Understanding other minds: Perspectives from developmental social neuroscience*, pp. 93–103, 2013.

[7] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *science*, vol. 315, no. 5812, pp. 619–619, 2007.

[8] A. Powers and S. Kiesler, "The advisor robot: tracing people's mental model from a robot's physical attributes," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pp. 218–225, 2006.

[9] J. Goetz, S. Kiesler, and A. Powers, "Matching robot appearance and behavior to tasks to improve human-robot cooperation," in *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, pp. 55–60, Ieee, 2003.

[10] S. M. Schaafsma, D. W. Pfaff, R. P. Spunt, and R. Adolphs, "Deconstructing and reconstructing theory of mind," *Trends in cognitive sciences*, vol. 19, no. 2, pp. 65–72, 2015.

[11] C. M. Heyes, "Theory of mind in nonhuman primates," *Behavioral and brain sciences*, vol. 21, no. 1, pp. 101–114, 1998.

[12] R. J. R. Blair, "Fine cuts of empathy and the amygdala: dissociable deficits in psychopathy and autism," *Quarterly journal of experimental psychology*, vol. 61, no. 1, pp. 157–170, 2008.

[13] S. Rossi, D. Conti, F. Garramone, G. Santangelo, M. Staffa, S. Varrasi, and A. Di Nuovo, "The role of personality factors and empathy in the acceptance and performance of a social robot for psychometric evaluations," *Robotics*, vol. 9, no. 2, p. 39, 2020.

[14] A. Rossi, P. Holthaus, G. Perugia, S. Moros, and M. Scheunemann, "Trust, acceptance and social cues in human–robot interaction (scrita)," 2021.

[15] S. H. Seo, D. Geiskkovitch, M. Nakane, C. King, and J. E. Young, "Poor thing! would you feel sorry for a simulated robot? a comparison of empathy toward a physical and a simulated robot," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 125–132, IEEE, 2015.

[16] Y. Suzuki, L. Galli, A. Ikeda, S. Itakura, and M. Kitazaki, "Measuring empathy for human and robot hand pain using electroencephalography," *Scientific reports*, vol. 5, no. 1, pp. 1–9, 2015.

[17] E. S. Cross, K. A. Riddoch, J. Pratts, S. Titone, B. Chaudhury, and R. Hortensius, "A neurocognitive investigation of the impact of socializing with a robot on empathy for pain," *Philosophical Transactions of the Royal Society B*, vol. 374, no. 1771, p. 20180034, 2019.

[18] J. Złotowski, H. Sumioka, S. Nishio, D. F. Glas, C. Bartneck, and H. Ishiguro, "Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy," *Paladyn, Journal of Behavioral Robotics*, vol. 7, no. 1, 2016.

[19] N. Spatola and O. A. Wudarczyk, "Ascribing emotions to robots: Explicit and implicit attribution of emotions and perceived robot anthropomorphism," *Computers in Human Behavior*, vol. 124, p. 106934, 2021.

[20] M. Blut, C. Wang, N. V. Wünderlich, and C. Brock, "Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other ai," *Journal of the Academy of Marketing Science*, vol. 49, no. 4, pp. 632–658, 2021.

[21] M. R. Fraune, "Our robots, our team: Robot anthropomorphism moderates group effects in human–robot teams," *Frontiers in Psychology*, vol. 11, p. 1275, 2020.

[22] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 33–42, 2020.

[23] H. Tan, D. Wang, and S. Sabanovic, "Projecting life onto robots: The effects of cultural factors and design type on multi-level evaluations of robot anthropomorphism," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 129–136, IEEE, 2018.

[24] E. Broadbent, V. Kumar, X. Li, J. Sollers 3rd, R. Q. Stafford, B. A. MacDonald, and D. M. Wegner, "Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality," *PloS one*, vol. 8, no. 8, p. e72589, 2013.

[25] F. Eyssel, D. Kuchenbrandt, F. Hegel, and L. De Ruiter, "Activating elicited agent knowledge: How robot and user features shape the perception of social robots," in *2012 IEEE RO-MAN: The 21st IEEE international symposium on robot and human interactive communication*, pp. 851–857, IEEE, 2012.

[26] C. R. Crowell, J. C. Deska, M. Villano, J. Zenk, and J. T. Roddy Jr, "Anthropomorphism of robots: study of appearance and agency," *JMIR human factors*, vol. 6, no. 2, p. e12629, 2019.

[27] K. Schaefer, "The perception and measurement of human-robot trust," 2013.

[28] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "How social robots influence people's trust in critical situations," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1020–1025, IEEE, 2020.

[29] A. Rossi, F. Garcia, A. C. Maya, K. Dautenhahn, K. L. Koay, M. L. Walters, and A. K. Pandey, "Investigating the effects of social interactive behaviours of a robot on people's trust during a navigation task," in *Annual Conference Towards Autonomous Robotic Systems*, pp. 349–361, Springer, 2019.

[30] J. Banks, "Theory of mind in social robots: replication of five established human tests," *International Journal of Social Robotics*, vol. 12, no. 2, pp. 403–414, 2020.

[31] D. Kahneman, *Thinking, fast and slow*. Macmillan, 2011.

[32] A. Waytz and M. I. Norton, "Botsourcing and outsourcing: Robot, british, chinese, and german workers are for thinking – not feeling – jobs.," *Emotion*, vol. 14, no. 2, p. 434, 2014.

[33] S. T. Fiske, A. J. Cuddy, and P. Glick, "Universal dimensions of social cognition: Warmth and competence," *Trends in cognitive sciences*, vol. 11, no. 2, pp. 77–83, 2007.

[34] K. Gray and D. M. Wegner, "Feeling robots and human zombies: Mind perception and the uncanny valley," *Cognition*, vol. 125, no. 1, pp. 125–130, 2012.

[35] E. Phillips, X. Zhao, D. Ullman, and B. F. Malle, "What is human-like? decomposing robots' human-like appearance using the anthropomorphic robot (abot) database," in *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, pp. 105–113, 2018.

[36] C. Canning and M. Scheutz, "Functional near-infrared spectroscopy in human-robot interaction," *Journal of Human-Robot Interaction*, vol. 2, no. 3, pp. 62–84, 2013.

[37] T. Himichi and M. Nomura, "Modulation of empathy in the left ventrolateral prefrontal cortex facilitates altruistic behavior: An fnirs study," *Journal of Integrative Neuroscience*, vol. 14, no. 02, pp. 207–222, 2015.

[38] U. Technologies, "Unity Real-Time Development Platform | 3D, 2D VR & AR Engine."

[39] U. Technologies, "Unity - Manual: Building your WebGL application," 2022.

[40] G. Andrews, G. S. Halford, K. M. Bunch, D. Bowden, and T. Jones, "Theory of mind and relational complexity," *Child development*, vol. 74, no. 5, pp. 1476–1499, 2003.

[41] B. Shneiderman and H. Hochheiser, "Universal usability as a stimulus to advanced interface design," *Behaviour & Information Technology*, vol. 20, no. 5, pp. 367–376, 2001.

[42] P. Sessa, F. Meconi, L. Castelli, and R. Dell'Acqua, "Taking one's time in feeling other-race pain: an event-related potential investigation on the time-course of cross-racial empathy," *Social cognitive and affective neuroscience*, vol. 9, no. 4, pp. 454–463, 2014.

[43] Y. Wang and S. Quadflieg, "In our own image? emotional and neural processing differences when observing human-human vs human-robot interactions," *Social cognitive and affective neuroscience*, vol. 10, no. 11, pp. 1515–1524, 2015.

[44] E. Wiese, G. Metta, and A. Wykowska, "Robots as intentional agents: using neuroscientific methods to make robots appear more social," *Frontiers in psychology*, vol. 8, p. 1663, 2017.