Theoretical Considerations for Social Learning between a Human Observer and a Robot Model

Theoretical Considerations for Social Learning between a Human Observer and a Robot Model

Boyoung Kim and Elizabeth Phillips
George Mason University

Robots are entering various domains of human societies, potentially unfolding more opportunities for people to perceive robots as social agents. We expect that having robots in proximity would create unique social learning situations where humans spontaneously observe and imitate robots' behaviors. At times, these occurrences of humans' imitating robot behaviors may result in a spread of unsafe or unethical behaviors among humans. For responsible robot designing, therefore, we argue that it is essential to understand physical and psychological triggers of social learning in robot design. Grounded in the existing literature of social learning and the uncanny valley theories, we discuss the human-likeness of robot appearance and affective responses associated with robot appearance as likely factors that either facilitate or deter social learning. We propose practical considerations for social learning and robot design.

INTRODUCTION

In the world where humans and robots coexist, there will be increasing opportunities for humans to observe and adopt behaviors of robots. When properly used, these unique social learning opportunities can promote prosocial attitudes and behaviors. However, when these opportunities are misused, it may set off a chain of unsafe or unethical attitudes and behaviors spread from a robot to a human and, subsequently, from a human to another human. For example, robots may not be perfectly immune to deliberate attempts of hacking or accidents of malfunctioning (Cerrudo & Apa, 2017), which could lead to situations where people learn antisocial behaviors from robots. Prior research has suggested that robot behavior is morally judged (Voiklis et al., 2016) and that robots can influence people to engage in (im)moral behaviors (Jackson & Williams, 2019). Thus, we propose that potential benefits and risks in these cases of social learning in Human-Robot Interaction (HRI) be closely monitored. To this end, it is necessary to examine physical and psychological factors of robots and their behaviors that may either promote or discourage people to model after them. Understanding people's overall tendency to imitate robots and specifying conditions under which this tendency becomes activated would help designing safe and ethical robots. In the following sections, we discuss the foundations of two theories that offer competing hypotheses for facilitating social learning in HRI.

SOCIAL LEARNING IN HUMAN-ROBOT INTERACTION

Social learning theorists propound that the observation of others plays an essential role for people to acquire new attitudes, skills, and behaviors (Bandura, 1986). Bandura in his social cognitive theory introduced four stages constituting the process of social learning (Bandura, 1986), which include attention, retention, reproduction, and motivation. The process begins with an observer paying attention to a model's behavior and its outcome, and proceeds with formulating a mental representation of the observed behavior and outcome. This mental representation needs to be retained in the observer's memory so that later the observer can reproduce the learned behavior. Whether the observer will reproduce the model's behavior is influenced by their motivation. For example, after a child observes and remembers their older sibling receiving a compliment from their parent for picking up trash, the child may be motivated to imitate the sibling's action. But, if the outcome of picking up trash was a criticism, the child would not be motivated to engage in the same behavior. This capacity to learn via observations of others enables people to acquire attitudes and behaviors that may help them deal with novel or uncertain situations, and without having to learn from direct prior experience.

Social learning researchers proposed that, from the very first stage of social learning, the attention stage, characteristics of the model have a significant influence on the observer's learning. For an observation of a model's behavior to facilitate a meaningful learning, it is necessary that the model's behavior first sufficiently holds the observer's attention. The extant literature has shown that the model's racial identity, biological sex or gender, belief, social power, and perceived competence, among other factors can influence whether the observer will mimic the model's behavior or not (Bandura, 1961; 1986; Schunk, 1987). We speculate that, if humans were to learn from robots via observations, their learning would also be affected by different characteristics of the robots and their ability to hold an observer's attention, as it is the case with human models.

Some HRI studies have demonstrated young children's tendency to imitate robot behavior via observations (Itakura et al., 2008; Sommer et al., 2020), but less research has been conducted to address whether and when adults would also spontaneously engage in learning through observations of robot behaviors. Further, the few studies that have targeted adult participants have offered inconsistent results. In one study, adult participants tended to mimic a robot player's decisions when the robot appeared to be good at earning monetary gain in a behavioral economic game (Zanatto et al., 2020); but, in another study, adult participants did not show more willingness to pick up trash after observing a robot picking up trash (Maeda et al., 2021). Also, whether the target population was children or adults, the primary goal of the previous studies was to test if humans can, and do successfully learn robot behaviors through observations. However, there has been little research examining the underlying mechanisms of such learning.

Many studies have illustrated that robot appearance is a powerful influencer of how humans will interact with robots. For instance, researchers have found that a human-like appearance can encourage humans to share responsibility with a robot (Broadbent et al., 2009), facilitate human-robot collaborations (Hinds et al., 2004); but also can elicit negative reactions from humans (Ho & MacDorman, 2010, 2017; Mori, 1970; Mori et al., 2012; Palomäki et al., 2018). Further, the human-like appearance of robots can impact perceptions of the types of roles robots are expected to fulfill (Goetz et al., 2003) and the types of moral decisions robots should make (Malle et al., 2015). Therefore, we expect that robot appearance can also influence the likelihood that people engage in observational learning from a robot.

In the current paper, we aim to extend the extant literature by introducing a research question that is grounded in the social cognitive theory (Bandura, 1986) and the uncanny valley theory (Mori, 1970; Mori et al., 2012). We discuss how physical similarities between humans and robots may affect social learning in a specific context of HRI where humans are the observer and robots are the model. Based on our review of the social cognitive (Bandura, 1986) and the uncanny valley theories (Mori, 1970), in this paper we first present our ideas about how learning via observations in these HRI contexts would be facilitated or interrupted by a robot's human-like physical appearance and affective responses associated with it; and then close with potential implications for robot design.

A ROBOT MODEL AND A HUMAN OBSERVER

Physical similarity between a human observer and a robot model

The perceived similarities between the observer and the model have been shown to promote social learning between humans. For example, in the classic Bobo doll experiment (Bandura et al., 1961), male children were more likely to mimic aggressive behaviors demonstrated by a male model than by a female model (and vice versa for female children), indicating the effect of the model-observer similarity in a biological sex or a gender. Given that the children in the experiment would have inferred the gender of the adult based upon physical cues, such as clothing, hairstyle, and voice, their findings at a broader level suggest the importance of the model-observer similarity of perceived physical appearance and outwardly observable characteristics.

Do these previous findings in the social learning literature suggest that people would be more likely to successfully learn from a robot when they observe a robot that is physically similar to themselves? Based on this model-observer similarity hypothesis (Bandura, 1986) in Human-Human Interaction (HHI), as stated above; a tentative answer to this question would be yes. The more a robot model appears similar to a human observer, the more likely a human observer would follow the robot model's behavior.

However, as the physical appearance of robots can vary vastly compared to that of humans, determining concrete bases for physical (dis)similarities between humans and robots may pose a challenge. For example, understanding how people identify whether a robot is male-gendered or female-gendered (or somewhere specifically on the spectrum of human gender) would be the subject of independent research in and of itself. Some researchers posited that the gender perception of a Pepper robot (Softbank Robotics) may diverge across different cultures (Søraa, 2017). Robot physical appearances can also include constellations of features that include both human-like and machine-like features as well as include a full range of features (from few to many), which vary in their realism and detail.

Moreover, determining a robot's perceived physical similarity with humans may be approached in multiple ways. First, a robot can be judged as looking similar to a human at a *holistic* level. This concept of the overall human-likeness of a robot has been essential for examining important research topics in HRI, such as the uncanny valley (Mori, 1970) and anthropomorphism towards robots (Duffy, 2003). Second, within the diverse spectrum of robot human-likeness, a robot's physical similarity to humans can be further decomposed into *specific* physical dimensions, or combinations of human-like features that systematically tend to co-occur together in real-world robots. As Phillips et al. (2018) found, human-like appearance in robots can be characterized by three human-like dimensions: the robots' Surface dimension (i.e., eyelashes, hair, skin, genderedness, nose, eyebrows, and apparel), the main components of its Body and Manipulators dimension (i.e., hands, arms, torso, fingers, legs), and its Facial dimension (e.g., face, eyes, head mouth). These dimensions are each uniquely and significantly predictive of people's overall perceptions of a robot's human-likeness (For further information see: www.abotdatabase.info/predictor). Thus, a robot can have a highly human-like face while having a highly mechanical body, which together give rise to a certain level of the overall human-likeness by a human perceiver.

In this paper, we focus on the human-likeness of a robot at the holistic level when considering the effect of physical similarity between a human observer and a robot model on social learning. Admittedly, this approach may not be able to offer explanations about which specific physical dimensions (or features or combinations thereof) of a robot will affect a human observer's learning of the robot's behavior. Future work on the effects of specific physical dimensions of robot human-likeness on social learning would be necessary. However, the holistic approach would provide foundations for making general and preliminary predictions for successful social learning in the context of HRI. Therefore, grounded in the model-observer similarity hypothesis (Bandura, 1986) in HHI, our initial prediction is that the more human-like a robot's overall physical appearance is, the more likely a human observer would imitate the robot model's behavior.

The human-likeness of a robot model and the problem of the uncanny valley

However, although robot human-likeness may facilitate social learning via model-observer similarity, high robot human-likeness poses a potential problem not discussed above known as the uncanny valley. The uncanny valley hypothesis

(Mori, 1970; Mori et al., 2012) posits that people would respond neutrally to robots that are low human-like and would increasingly respond positively towards robots as they resemble humans more. However, when robots resemble humans highly but not perfectly, people would suddenly exhibit strongly negative responses to these robots. After this interim phase dubbed the uncanny valley passes, robots that are almost perfectly human-like (e.g., androids) would elicit the most positive responses (See Figure 1).

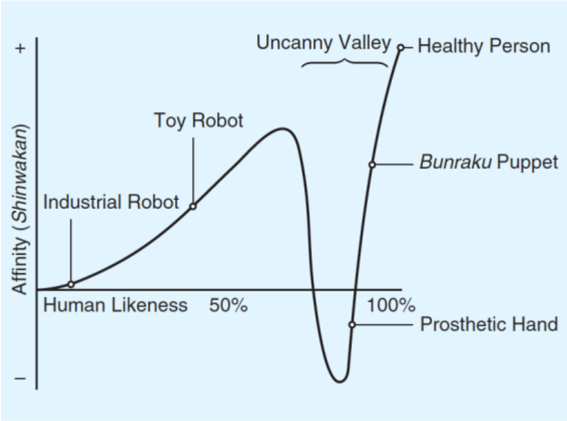

Figure 1. A visual depiction of the uncanny valley proposed by Mori (1970; Borrowed from Mori, MacDorman, & Kageki, 2012).

Drawing from the uncanny valley hypothesis, our prediction about human-likeness as a facilitator of social learning may not hold true in the context of HRI. To illustrate, when both the observer and the model roles were served by humans, a higher model-observer similarity would predict more success in learning (Bandura, 1986; Schunk, 1987). However, when the human was the observer and the robot was the model, the uncanny valley hypothesis suggests that a higher physical similarity (in this case, the human-likeness) would not necessarily predict a linearly positive relationship. When highly human-like robots that fall in the uncanny valley serve in the role of a model, would the robots' resemblance with humans facilitate the human observer's learning or, instead, would the negative responses elicited by those uncanny robots interfere with the observer's learning?

Many studies that investigated the relationships between robots' human-likeness and human responses to those agents showed their perceptions of the agents being high in creepiness, eeriness, and scariness and low in trustworthiness and likability (Abubshait et al., 2017; Ho & MacDorman, 2017; Mathur et al., 2020; Palomäki et al., 2018). Little research, on the other hand, has examined how uncanniness of highly human-like (and potentially uncanny) robots may interact with observational learning when a robot is a model and a human is an observer. Some research has shown that robots' human-like appearance can influence not only perceptions of robots as uncanny or creepy but also behavioral outcomes like whether such robots should be trusted. In these studies, participants were asked to make decisions on the amount of monetary endowment they would like to make to robots with varying degrees of physical human-likeness (Abubshait et al., 2017). The participants endowed less money to highly human-like robots that elicited uncanny responses compared to low human-like robots. If these findings were to be applicable to social learning settings in HRI, we would find an interference effect, as opposed to a facilitation effect for social learning, for highly human-like robots, especially if those robots trigger negative responses from people.

TENTATIVE HYPOTHESES

So far in this paper, our views on how a human observer would respond to a robot model in the context of social learning have mainly focused on robots that are highly human-like but not sufficiently human-like to bypass the uncanny valley. To prepare an empirical study for testing these ideas, we also attended to another category of robots that may present a contrast to uncanny robots, thereby offering different insights. Table 1 shows comparisons between these two groups of robots on physical similarity with humans and predicted resulting affective responses. Specifically, low human-like robots represent little to not at all human-like robots (e.g., Jibo), while high human-like robots represent the robots positioned along the deep curve of the uncanny valley (e.g., BINA48). These two categories of robots also diverge in terms of people's responses towards them. According to Mori et al. (2012), whereas low human-like robots are expected to induce rather neutral or ambivalent reactions from people, high human-like robots are expected to induce negative reactions.

Table 1. Predicted relationships between robot appearance (low vs. high human-like) and affective responses. The table depicts the Jibo robot developed by Cynthia Brezeal and the BINA48 robot developed by Hanson Robotics.

	Low Human-like Robots	**High Human-like (Uncanny) Robots**
	Jibo	**BINA48**
Physical Similarity with Humans	Low	High
Affective Response	Neutral	Negative

When combining social learning theory and the model-observer similarity hypothesis with the uncanny valley hypothesis, we are thus left with two opposing hypotheses about how likely it is that a human observer would successfully learn from a robot model through observation.

Hypothesis 1. If the model-observer similarity hypothesis in the HHI literature transferred to social learning settings in HRI, high human-like robots would seize more attention than low human-like robots would. Thus, a human observer would learn more from a robot model when the robot had a highly human-like appearance compared to when the robot had a little human-like appearance.

Hypothesis 2. On the contrary, if the uncanny valley hypothesis in the HRI literature transferred to social learning settings in HRI, high human-like robots would elicit more negative affective responses from people compared to low human-like robots. Thus, a human observer would learn more from a robot model when the robot had a little human-like appearance compared to when the robot had a highly human-like appearance.

Essentially, these two hypotheses suggest differential mechanisms by which social learning would be promoted/diminished by robot appearance. Hypothesis 1 posits that robot appearance impacts social learning through model-observer similarity and its ability to capture the observer's attention, while Hypothesis 2 proposes that robot appearance impacts social learning via perceptions of the robots (un)creepiness, etc. Thus, additional research is needed to determine which of these hypotheses and associated underlying psychological mechanisms best predicts social learning from robots in HRI, and how social learning can be enhanced or diminished by robot appearance.

CONCLUSIONS AND IMPLICATIONS FOR DESIGN

Humans are predisposed to refer to other social agents to establish attitudes and choose proper behaviors. As humans share the social space with more robots in the future, it is possible (and perhaps natural) for social learning to occur, specifically, in the form of a human as an observer and a robot as a model. We expect that these occurrences of social learning may mostly lead to harmless or beneficial outcomes but sometimes may cause harm to humans when the adopted behaviors are unsafe or unethical. We know from prior research for instance, that robots can persuade people to engage in morally charged behaviors (Jackson & Williams, 2019). We argue, therefore, that robot designers should take heed of potential triggers of social learning, as doing so can both facilitate the modeling of prosocial behaviors and diminish anti-social behaviors, especially when robots are deployed in large-group or public settings in which social learning would be important, like airports, shopping malls, or schools.

It is also possible that observational learning from robots in these settings may have a ripple effect for continued observational learning from humans. The kernel of the ripple effect begins with a human learning from a model robot, but then becomes a model for other humans and so forth. In the present paper, we attempted to bridge the HHI literature in social learning and the HRI literature to draw predictions and hypotheses about how physical human-likeness and the related psychological characteristics of robots may influence the process of social learning. We believe that continued efforts to examine social learning in the context of HRI would be crucial for robot design, as social learning and the uncanny valley theories suggest competing hypotheses for how robot appearance could influence social learning from robots. Understanding the mechanisms underlying social learning in HRI can provide bases for creating environments that promote positive or desired behaviors in everyday life as well as more purposeful skill acquisition settings like training applications. Thus, we propose the following considerations for designing robots based on our review:

- When considering the effects of human-likeness of robot appearance on social learning, we recommend taking both holistic and specific physical dimension-based approaches to defining human-likeness.
- When considering the effects of robot human-like appearance on social learning, we recommend identifying the possible influences of the uncanny valley and their resulting effects on social learning.
- To design a robot that induces people to engage in good deeds while avoiding bad deeds, considerations of both physical human-likeness and affective responses associated with varying degrees of human-likeness are necessary.

ACKNOWLEDGEMENTS

This work was supported in part by NSF grant IIS-1909847.

REFERENCES

Abubshait, A., Momen, A., & Wiese, E. (2017). Seeing human: Do individual differences modulate the Uncanny Valley? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 870–874.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory* (pp. xiii, 617). Prentice-Hall, Inc.

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *The Journal of Abnormal and Social Psychology*, *63*(3), 575–582.

Broadbent, E., Stafford, R., & MacDonald, B. (2009). Acceptance of Healthcare Robots for the Older Population: Review and Future Directions. *International Journal of Social Robotics*, *1*(4), 319.

Cerrudo, C., & Apa, L. (2017). Hacking Robots Before Skynet. *IOActive Website*, 1–17.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3), 177–190.

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, 55–60.

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose Job Is It Anyway? A Study of Human-Robot Interaction in a Collaborative Task. *Human-Computer Interaction*, *19*(1/2), 151–181.

Ho, C.-C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior*, *26*(6), 1508–1518.

Ho, C.-C., & MacDorman, K. F. (2017). Measuring the Uncanny Valley Effect. *International Journal of Social Robotics*, *9*(1), 129–139.

Itakura, S., Ishida, H., Kanda, T., Shimada, Y., Ishiguro, H., & Lee, K. (2008). How to Build an Intentional Android: Infants' Imitation of a Robot's Goal-Directed Actions. *Infancy*, *13*(5), 519–532.

Jackson, R. B., & Williams, T. (2019). Language-Capable Robots may Inadvertently Weaken Human Moral Norms. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 401–410.

Maeda, R., Bršćić, D., & Kanda, T. (2021). Influencing Moral Behavior Through Mere Observation of Robot Work: Video-based Survey on Littering Behavior. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 83–91.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 117–124.

Mathur, M. B., Reichling, D. B., Lunardini, F., Geminiani, A., Antonietti, A., Ruijten, P. A. M., Levitan, C. A., Nave, G., Manfredi, D., Bessette-Symons, B., Szuts, A., & Aczel, B. (2020). Uncanny but not confusing: Multisite study of perceptual category confusion in the Uncanny Valley. *Computers in Human Behavior*, *103*, 21–30.

Mori, M. (1970). Bukimi no tani [the uncanny valley]. *Energy*, *7*, 33–35.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics Automation Magazine*, *19*(2), 98–100.

Palomäki, J., Kunnari, A., Drosinou, M., Koverola, M., Lehtonen, N., Halonen, J., Repo, M., & Laakasuo, M. (2018). Evaluating the replicability of the uncanny valley effect. *Heliyon*, *4*(11), e00939.

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). What is Human-like? Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 105–113.

Schunk, D. H. (1987). Peer Models and Children's Behavioral Change. *Review of Educational Research*, *57*(2), 149–174.

Sommer, K., Davidson, R., Armitage, K. L., Slaughter, V., Wiles, J., & Nielsen, M. (2020). Preschool children overimitate robots, but do so less than they overimitate humans. *Journal of Experimental Child Psychology*, *191*, 104702.

Søraa, R. A. (2017). Mechanical genders: How do humans gender robots? *Gender, Technology and Development*, *21*(1–2), 99–115.

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). Moral judgments of human vs. Robot agents. *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 775–780.

Zanatto, D., Patacchiola, M., Goslin, J., Thill, S., & Cangelosi, A. (2020). Do Humans Imitate Robots?: An Investigation of Strategic Social Learning in Human-Robot Interaction. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 449–457.