# When Robots Self-Disclosure Personal Information, Do Users Reciprocate?

**Jessica K. Barfield**
University of Tennessee- Knoxville
jbarfiel@vols.utk.edu

**Abstract**

Human-robot interaction consists of a rich set of behaviors between humans and robots often requiring the exchange of personal and sensitive information between them. From a conceptual framework this paper discusses whether a robot who self-discloses personal information when conversing with a user will prompt the user to reciprocate and self-disclose personal and sensitive information to the robot. Additionally, the paper discusses various factors which may influence whether self-disclosure of personal information between human and robot occurs and briefly discusses aspects of a conceptual representational system necessary for HRI enabling the robot to self-disclose to a user.

## Introduction

Robots are increasingly entering society interacting with humans in contexts which require sophisticated social skills. One issue of importance for human-robot interaction (HRI) is whether humans should be required to disclose personal, private, or sensitive information to a robot when experienced in social contexts (Barfield, 2023; Kumazaki et al. 2022). For example, care robots may require that people disclose personal health information to a robot, robo-advisors may require sensitive financial information to be disclosed, and robots serving as a behavioral counselor may require sensitive and potentially embarrassing information about relationships and mental health to be disclosed. From research in psychology, one factor which may determine whether an individual discloses personal information to another, is whether the disclosure is in response to or prompted by a personal disclosure first made by the other person (Barfield, 2021; Cozby, 1973; Eyssel, et al. 2017; Mou et al. 2023; Neerinex et al. 2022; Zhu et al. 2023). In fact, we have a tendency to disclose personal information, even to a stranger, if that person has first disclosed personal information to us. In this paper, drawing from the above discussion on self-disclosure from the psychology literature, I explore whether the act of self-discloser of personal and sensitive information to a robot is dependent on whether the robot first disclosed personal information to the user. In this paper I do not argue that human disclosure of personal and sensitive information to a robot is the default condition after robot disclosure, but instead I propose a host of factors such as the context of the communication, amount of trust in the robot, anthropomorphism of the robot, and others are important factors to consider.

## Self-Disclosure with Robots

In a study using the Nao robot, Eyssel, Wullenkord, and Nitsch (2017) described self-disclosure as the act of revealing personal information, thoughts, and feelings to another party. Further, they commented that self-disclosure is an important determinant of liking in relationships and is central to the development of close relationships among humans. From this follows a testable prediction for HRI that we should like a robot if we have disclosed personal information to the robot. Additionally, so too did Neerinerx et al. (2022) state that the act of self-disclosure is an important aspect of HRI (see also, Xiaoling, et al. 2023). Using a field study, they investigated the human tendency to self-disclose to the robot Pepper when visitors to a museum were allowed to interact and converse with Pepper. Based on observations and survey results, when visitors were questioned by Pepper, that is, when Pepper initiated the conversation, they disclosed more of their attitudes and opinions concerning the museum than about other topic categories suggesting that robot disclosure may influence the type of information disclosure by the user. In another study, Eyssel and colleagues (2017) manipulated whether the Nao robot disclosed personal information or whether Nao asked personal questions to a human partner to elicit information. Interestingly, in contrast to the above prediction that people would like a robot more if they disclosed to it, their results found no statistically significant effects of self-disclosure for measures of robot likability or HRI quality. Furthermore, their results indicated that being in the role of a passive listener versus actively conversing with the Nao robot affected HRI more than

the actual content of the verbal exchange. However, a related study on whether user disclosure of information resulted in liking the robot produced somewhat different results. Mou et al. (2023) found that reciprocal self-disclosure from a robot increased liking for conditions of intimate self-disclosure but decreased liking in non-intimate self-disclosure. Barfield (2021) also investigated self-disclosure of personal and sensitive information to robots of different appearances and found that robots with a friendly facial expression were more likely to prompt self-disclosure.

In a different context for self-disclosure, Kumazaki and colleagues (2022) studied individuals with autism spectrum disorders (ASD) and used sentence completion tests (SCT) to investigate self-disclosure for individuals with ASD. They compared the difference in disclosure statements to an android robot and a human interviewer. In contrast to the results reported by Barfield (2021) their results suggested that the android robot promoted more self-disclosure, especially about a negative topic compared to a human interviewer. Considering the two studies together clearly indicates that additional research is necessary to more fully understand the conditions under which humans might disclose personal information to robots. Focusing on an elderly population, Noguchi, Kamide, and Tanaka (2018) explored various characteristics of robots which could encourage self-disclosure by seniors. They found that if a robot engaged in human-like behavior, this could encourage elderly people's self-disclosure about the experience of loss. Additionally, considering gender differences and self-disclosure, Uchida et al. (2020) hypothesized that compared to men, women would be less likely to discriminate between robots and humans as listeners for their self-disclosure or in the amount of self-disclosure. Finally, van Straten et al. (2022) commented that while scholars have focused on the role of self-disclosure in the context of child-robot interaction, little is known as to how the effects of a robot's self-disclosure varies by the information the robot shares. They investigated how children perceived the Nao robot when the robot engaged in personal self-disclosure versus factual self-descriptions and when the robot did or did not ask questions. Among others, their results indicated that the robot's question-asking increased children's trust in the robot. Summarizing the above studies, factors that affect whether people disclose personal and sensitive information to robots may depend on gender differences in participants, the type of information to be shared, the facial features of a robot, the extent of self-disclosure by another party, and whether there is a passive versus active listener. I next discus the conceptual representation space between human and robot for personal information disclosure.

## The Representation Space of Self-Disclosure

When considering HRI, robots are often provided a cognitive map or schema to represent and support the demand characteristics of the social interaction (Yan, Weber, and Wermter, 2012). Considering the topic of this paper, this could be a set of instructions or rules guiding the extent of self-disclosure of personal information from the robot to the user based on responses received from the user. However, as indicated by Jung et al. (2007), endowing human-like characteristics to a robot is a difficult problem which requires representations which contain the richness and diversity of HRI. On this point, Bobu (2023) indicated that to determine the human's representation of what matters in a task requires methods that rely on data sets of human and robot behavior; for example, in the context of this paper, this could be the extent to which each party discloses information to each other. Additionally, as discussed by Bobu (2023) we should treat humans and robots as active participants in their interactions with each other and not as static sources of data a comment which clearly applies to human-robot self-disclosure. In an earlier work on robot representations of a problem space, Belouaer, Bouzid, and Mouaddib (2010) commented that HRI requires that robots be endowed with spatial representation and/or a reasoning system, but with relevance to self-disclosure, most robots are not programmed to consider fuzzy relations given by linguistic variables in humans language (see also, Zivkovic et al. 2008). Of course, since this paper was published, much progress has been made in developing algorithms to represent the cognitive demands of HRI. For example, to address a robot's representation of context for HRI, Zachary et. al. (2015) discussed CARIL (Context-Augmented Robotic Interaction Layer) which they described as a human-robot interaction system that leveraged cognitive representations of shared context as a basis for HRI. Such a system seems appropriate to model human-robot self-disclosure performance in HRI. Clearly, among roboticists there has been much research on providing robots with a cognitive representation or framework to guide a robot's behavior. One example is provided by Arnold, Kasenberg, and Scheutz (2021) who described a cognitive architecture for supporting HRI with algorithms for generating explanations relevant for HRI. In their system, explanations were used to track the robot's actual decision-making and behavior, which were determined by normative principles the robot used to justify its actions. Of relevance for self-disclosure among humans and robots, they concluded that for HRI normative principles strongly guide what information is shared between individuals and thus should be the focus of a representational system for robots in HRI.

Additionally, Chella et al. (2008) proposed an Entertainment Humanoid Robot model based on the technique of Latent Semantic Analysis, to produce an emotional behavior in a robot's interaction with humans (see also Chella et al. 2004). Their approach allows the coding of a words semantics by specific statistical computations applied to a large set of text; such an approach seems particularly relevant for self-disclosure between humans and robots. Further, Kobayashi, Murata, and Inamura (2022) discussed a data-driven framework for analyzing physical human-robot interaction. The model was discussed as being critical for elaborating human understanding and/or robot control during HRI. And Lemaignan, et al. (2017) commented that HRI is in many ways dynamic, involving unknown environments that were not originally designed for robots consisting of a broad variety of situations consisting of rich semantics and which may include divergent mental models between human and robots for a given interaction. From the above brief discussion and as a summary, the following figure outlines the idea that the self-disclosure of personal and sensitive information to a robot is based on a complex relationship between characteristics of the user and robot; and for human-robot self-disclosure requires a representational system that takes human and robot characteristics into account.

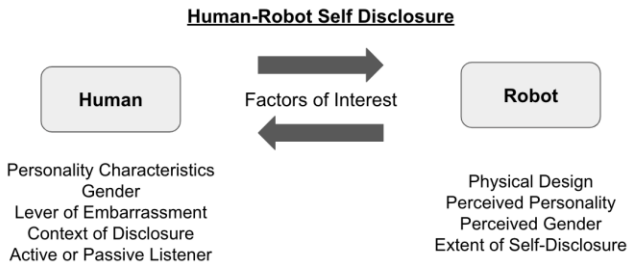


Figure 1: Proposed Factors Influencing Human-Robot Interaction and Self-Disclosure of Personal Information

In the above figure various characteristics of the person interacting with a robot may determine whether they will self-disclose personal information to a robot, such as their personal characteristics (e.g., extrovert versus introvert personality), their gender, age, race, and ethnicity. In addition, the design of the robot may be a factor in whether an individual will self-disclose to a robot, such as the degree to which a robot is anthropomorphized, its facial features, voice characteristics, and other design features of the robot. From this, it is clear that for the self-disclosure of personal information between humans and robots the representational system enabling conceptualization of the HRI is of primary importance.

## References


Arnold, T., Kasenberg, D., and Scheutz, M., Explaining in Time: Meeting Interactive Standards of Explanation for Robotic Systems, ACM Transactions on Human-Robot Interaction, Vol. 10 (3), 1-23, 2021.

Barfield, J. K., Evaluating the Self-Disclosure of Personal Information to AI Enabled Technology (book chapter), in Seungahn, Nah (Ed.), Research Handbook on Artificial Intelligence and Communication, Edward Elgar Press. Forthcoming.

Barfield, J. K., Self-Disclosure of Personal Information, Robot Appearance, and Robot Trustworthiness, 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 67-72, 2021.

Belouaer, L., Bouzid, M., and Mouaddib, A. I., A Spatial Ontology for Human-Robot Interaction, 7th International Conference on Informatics in Control, Automation and Robotics, Vol. 1, 154-159, 2010.

Bobu, A., Aligning Robot and Human Representations, NYU Colloquium. https://cs.nyu.edu/dynamic/news/colloquium/1303/. Accessed: 2023-03-30.

Chella, A., Infantino, I., and Macaluso, I., Conceptual Spaces and Robotic Emotions

IEEE International Conference on Systems, Man and Cybernetics, Vol. 1-7, 144-149, 2004,

Chella, A., Pilato, G., Sorbello, R., Vassallo, G., Cinquergrani, F., and Anzalone, S. M., An Emphatic Humanoid Robot with Emotional Latent Semantic Behavior, 1st International Conference on Simulation, Modeling and Programming for Autonomous Robots, 234-245, 2008.

Cozby, P. C., Self-disclosure: A Literature Review. Psychological Bulletin, Vol. 79(2), 73–91, 1973.

Eyssel, F., Wullenkord, R., and Nitsch, V., The Role of Self-Disclosure in Human-Robot Interaction, 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 922-927, 2017.

Jung, Y., Choi, Y., Park, H., Shin, W., and Myaeng, S. H., Integrating Robot Task Scripts With a Cognitive Architecture for Cognitive Human-Robot Interactions, IEEE International Conference on Information Reuse and Integration (IRI-2007), 152-157, 2007.

Kobayashi, T., Murata, S., and Inamura, T., Latent Representation in Human-Robot Interaction With Explicit Consideration of Periodic Dynamics, IEEE Transactions on Human-Machine Systems, Vol. 52 (5), 928-940, 2022.

Kumazaki, H., Muramatsu, T., Yoshikawa, Y., Takata, K., Ishiguro, H., and Mimura, M., Android Robot Promotes Disclosure of Negative Narratives by Individuals With Autism Spectrum Disorders, Frontiers in Psychiatry, Vol. 13, 2022.

Lemaignan, S., Warnier, M., Sisbot, E. A., Clodic, A., and Alami, R., Artificial Cognition for Social Human-Robot Interaction: An Implementation, Artificial Intelligence, Vol. 247, 45-69, 2017.

Mou, Y., Zhang, L., Wu, Y., Pan, S., and Ye, X. Y., Does Self-Disclosing to a Robot Induce Liking for the Robot? Testing the Disclosure and Liking Hypotheses in Human-

Robot Interaction, International Journal of Human–Computer Interaction, 2023.

Neerincx, A., Edens, C., Broz, F., Li, Y., and Neerincx, M., Self-Disclosure to a Robot "In-the-Wild": Category, Human Personality and Robot Identity, 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) - Social, Asocial, and Antisocial Robots, 584-591, 2022.

Noguchi, Y., Kamide, H., and Tanaka, F., Effects of the Self-Disclosure of Elderly People by Using a Robot with Intermediates Remote Communication, 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 612-617, 2018.

Uchida, T., Takahashi, H., Ban, M., Shimaya, J., Minato, T., and Ogawa, K., Japanese Young Women Did not Discriminate between Robots and Humans as Listeners for Their Self-Disclosure -Pilot Study- Multimodal Technologies and Interaction, Vol. 4 (3), 2020.

van Straten, C. L., Peter, J., Kuhne, R., and Barco, A., On Sharing and Caring: Investigating the Effects of a Robot's Self-disclosure and Question-Asking on Children's Robot Perceptions and Child-Robot Relationship Formation, Computers in Human Behavior, pg. 107135, 2022.

Yan, W., Weber, C., and Wermter, S., A Neural Approach for Robot Navigation Based on Cognitive Map Learning, The 2012 International Joint Conference on Neural Networks (IJCNN), 1-8, 2012.

Zachary, W., Johnson, M., Hoffman, R., Thomas, T., Rosoff, A., and Santarelli, T., A Context-Based Approach to Robot-Human Interaction, 6th International Conference on Applied Human Factors and Ergonomics (AHFE), 1052-1059, 2015.

Zhu, X., Liang, W., Xv, W., and Wang, Y., The Key Strategies for Increasing Users' Intention of Self-Disclosure in Human-Robot Interaction Through Robotic Appearance Design, SHS Web of Conferences, Vol.165, pg. 01012, 2023.

Zivkovic, Z., Booij, O., KrOse, B., Topp, E. A., and Christensen, H. I., From Sensors to Human Spatial Concepts: An Annotated Data Set, IEEE Transactions on Robotics, Vol. 24 (2), 501-505, 2008.