How does a robot's social credibility relate to its perceived trustworthiness?

Patrick Holthaus
School of Physics, Engineering, and Computer Science
University of Hertfordshire
Hatfield, United Kingdom
p.holthaus@herts.ac.uk

***Index Terms*—Human-robot interaction; social credibility; trustworthiness**

I. Introduction

This position paper aims to highlight and discuss the role of a robot's *social credibility* in interaction with humans. In particular, I want to explore a potential relation between social credibility and a robot's acceptability and ultimately its trustworthiness. I thereby also review and expand the notion of social credibility as a measure of how well the robot obeys social norms during interaction [1] with the concept of conscious acknowledgement.

II. Social robot acceptability

Humans interaction comprises social signals that nonverbal exchange theories often argue to enrich the communication channel with additional information [2]. Likewise, robots are often designed in a sociable way and programmed to exhibit social competence with the aim to improve the quality and effectiveness of an interaction [3]. A wide range of experiments in the research field of human-robot interaction (HRI) evaluate particular elements or combinations of robot appearances and designs or verbal and nonverbal behaviours with regard to the human's social perception of the robot.

Many different factors affecting people's social perception of a robot have already been identified, such as the naturalness of a robot's movements [4], expressiveness and vulnerability of a robot [5], emotions [6], and proximity [7]. Social behaviours can provide a robot with the ability to establish and maintain social relationships by using natural cues, and expressing and perceiving emotions [8]. To be acceptable by humans, such social behaviour must be meaningful and congruent [9], appropriate to the social role that it is expected to fulfil [10], and continuously provides appropriate signals for the entire duration of an interaction [11].

Typically, the acceptability of a robot's social functions is verified using questionnaire scales like *Godspeed* [12] or the *Robot Social Attribute Scale* (RoSAS) [13] that *implicitly* measure the participants' perception of the robot in different experimental conditions. That is, participants normally rate general robot attributes like reliability, competency, happiness, or scariness. Qualitative evaluation, open questions, and objective measures like participant compliance with robot requests help to further reason about user ratings and preference. As a result, appropriate robot sociability, i.e. design and behaviours that lead to the robot's acceptability, are usually acknowledged subconsciously by human participants and exact reasons are often interpreted using additional data.

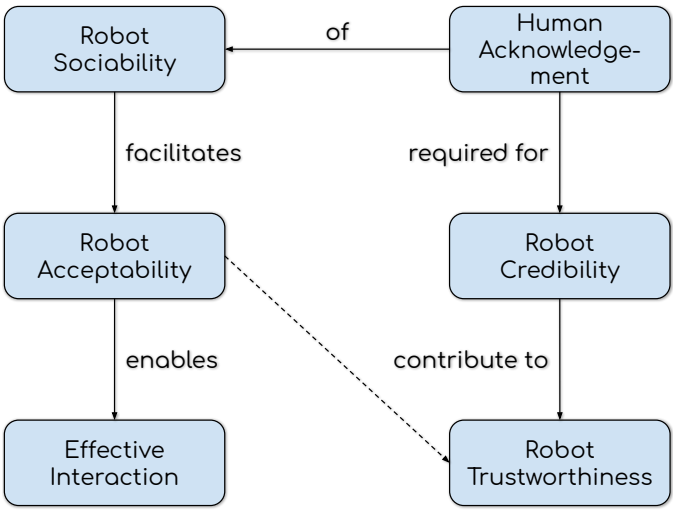

Fig. 1. Robot sociability: How a human's conscious and unconscious acknowledgement might lead to acceptability, credibility and trustworthiness

III. Trust in social robots

Trust is one of the fundamental factors for a successful cooperation between humans and robots [14]–[16]. The research field of HRI employs and develops several definitions of trust that originate in psychology and are then adapted to interaction with robots. Most of these definitions (e.g. [17]) include cognitive and affective factors that contribute people's to assessment of an robots reliability by its actions and behaviours. On this basis, people can form an emotional connection with robots on the assumption that both (human and robot) are positively invested in the success of the interaction and relationship [18], [19]. Consequently, the social acceptability of robot behaviours can heavily influence a robot's perceived trustworthiness.

IV. Social robot credibility

A robot's *social credibility* has been introduced as a measure of "how well it obeys the social norms relevant to its environ-

ment" [1]. As such, the term is closely related to the notion of socially acceptable behaviour but not entirely identical. Behaviours can be socially credible but not acceptable, e.g. when a robot is impolite or interrupting a conversation. At the same time, they can also be acceptable but not credible, e.g. when a robot is pretending to empathise with human emotions. The concept of social credibility aims to quantify how believable and authentic a robot's social behaviours are. It thereby explicitly considers the context of existing social norms where robot behaviour can be credible by staying within the range of behaviours that would be expected from a human interaction partner.

The definition further implies that a robot's social credibility is always evaluated with regard to the environment and robot itself. That is, the credibility of a behaviour is tied to a situation with one specific robot in a particular environment as expectations on normative behaviour might be in a different in other situations. Much like behaviours that are socially not acceptable, socially incredible robots are more likely to be perceived negatively by people during an interaction. For example, a robot might be perceived as having less authority [20] when acting as a safety monitor.

V. Influence of credibility on trustworthiness

In addition to the above, it is noteworthy that if a certain robot behaviour is supposed to be credible, a human needs to acknowledge it as deliberately exhibited by the robot with the specific purpose to facilitate social interaction. Social credibility as the humans conscious acknowledgement of a robot's sociability might be an important factor that contributes to a robot's trustworthiness. If we are able to capture whether users recognise or not recognise a robot's social actions as such we could further develop the conscious part of a humans mental model about a robot that affect its perceived trustworthiness. At the same time, the acceptability of robot behaviours contributes to the person's mental model of the robot, albeit to the subconscious part, cf. Fig. 1.

The credibility of a robot's social behaviour might also have an important role when the human's trust in the robot changes or is lost. Robot actions that highlight the positive investment in the interaction by demonstrating its engagement in social behaviour that do not serve any obvious other function have the potential to be used as repair mechanisms. Explicit social signals like a communicative gaze, blinking, facial expression, gesture, or a colloquial utterance might have a positive effect in such situations even if they are interrupting the otherwise acceptable interaction.

It is not yet entirely clear how to quantify a robot's social credibility. I argue that we need to capture the human's conscious acknowledgement of the robot's sociability to successfully measure social credibility and draw better conclusion about a robot's perceived trustworthiness. It might therefore be worthwhile to develop and adopt subjective and objective measurement standards that allow conclusion about a robot's social credibility as a complementary method to sociability in HRI experiments.

References

[1] C. Menon and P. Holthaus, "Does a Loss of Social Credibility Impact Robot Safety? Balancing social and safety behaviours of assistive robots," in *International Conference on Performance, Safety and Robustness in Complex Systems and Applications (PESARO)*. ARIA, 2019, pp. 18–24.

[2] M. L. Patterson, "A Sequential Functional Model of Nonverbal Exchange," *Psychological Review*, vol. 89, no. 3, pp. 231–249, 1982.

[3] A. Esposito and L. C. Jain, "Modeling Social Signals," in *Toward Robotic Socially Believable Behaving Systems-Volume II*. Springer, 2016, vol. 106.

[4] C. Lichtenthaler, T. Lorenzy, and A. Kirsch, "Influence of legibility on perceived safety in a virtual human-robot path crossing task," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2012, pp. 676–681.

[5] N. Martelaro, V. C. Nneji, W. Ju, and P. Hinds, "Tell me more designing hri to encourage more trust, disclosure, and companionship," in *ACM/IEEE International Conerence on Human-Robot Interaction (HRI)*, 2016, pp. 181–188.

[6] M. Saerbeck and C. Bartneck, "Perception of affect elicited by robot motion," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010, pp. 53–60.

[7] K. L. Koay, D. S. Syrdal, M. L. Walters, and K. Dautenhahn, "Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 2007, pp. 564–569.

[8] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 143–166, 2003.

[9] F. Hegel, S. Gieselmann, A. Peters, P. Holthaus, and B. Wrede, "Towards a Typology of Meaningful Signals and Cues in Social Robotics," in *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Atlanta, Georgia, 2011, pp. 72–78.

[10] K. L. Koay, D. S. Syrdal, M. Ashgari-Oskoei, M. L. Walters, and K. Dautenhahn, "Social roles and baseline proxemic preferences for a domestic service robot," *International Journal of Social Robotics*, vol. 6, no. 4, pp. 469–488, 2014.

[11] P. Holthaus and S. Wachsmuth, "It was a Pleasure Meeting You - Towards a Holistic Model of Human-Robot Encounters," *International Journal of Social Robotics*, 2021.

[12] C. Bartneck, E. Croft, and D. Kulic, "Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, p. 71–81, 2009.

[13] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, "The Robotic Social Attributes Scale (RoSAS): Development and Validation," in *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2017, pp. 254–262.

[14] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors: The Journal of Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, 2011.

[15] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Towards safe and trustworthy social robots: Ethical challenges and practical issues," in *Social Robotics*, ser. Lecture Notes in Computer Science, vol. 9388. Cham: Springer International Publishing, 2015, pp. 584–593.

[16] A. Rossi, K. Dautenhahn, K. L. Koay, and M. L. Walters, "A study on how the timing and magnitude of robot errors may influence people trust of robots in an emergency scenario," in *International Conference on Social Robotics (ICSR)*, ser. Lecture Notes in Computer Science, vol. 10652. Cham: Springer International Publishing, 2017, pp. 42–52.

[17] J. D. Lewis and A. Weigert, "Trust as a social reality." *Social Forces*, vol. 63, no. 4, p. 967–985, 1985.

[18] D. J. McAllister, "Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations," *Academy of Management Journal*, vol. 38, no. 1, pp. 24–59, 1995.

[19] E. J. de Visser, M. M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams." *International Journal of Social Robotics*, p. 459–478, 2020.

[20] P. Holthaus, C. Menon, and F. Amirabdollahian, "How a Robot's Social Credibility Affects Safety Performance," in *International Conference on Social Robotics (ICSR)*, ser. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, vol. 11876, pp. 740–749.