Social Injustice, Group Membership and Epistemic Trust in Robots

Stedtler, Samantha

2024

Document Version:
Publisher's PDF, also known as Version of record

Citation for published version (APA):
Stedtler, S. (2024). *Social Injustice, Group Membership and Epistemic Trust in Robots*. 1-4. Paper presented at Robo-Identity Workshop 3, HRI'24: Robo-Identity: Designing for Identity in the Shared World.

Total number of authors:
1

Social Injustice, Group Membership and Epistemic Trust in Robots

Samantha Stedtler
samantha.stedtler@lucs.lu.se
Lund University
Lund, Sweden

ABSTRACT

This paper focuses on the intersection of social dynamics, knowledge production, and the implementation of identity features in robotic platforms. Concentrating on the concept of epistemic injustice, particularly in the context of trust, the study highlights the potential for justified distrust based on experienced injustice. The incorporation of social identity features, such as gender, signals group membership and can instigate trust based on assumed shared experiences of oppression. However, I argue that this can lead to a form of deception, especially for oppressed groups, as robots lack the embodied experience of living through inequality. The paper explores the risk of companies, largely composed of privileged individuals, potentially exploiting this trust. Three key suggestions for designing diverse robot identities are proposed: 1) Users should have control over robot features, emphasizing the non-fixed nature of robot identity 2) Diverse identity features should be combined to avoid overtrust and stereotyping 3) Robots should encourage users to rely on their own knowledge, fostering self-trust and preventing the perpetuation of privileged perspectives.

CCS CONCEPTS

• **Social and professional topics** → **User characteristics**; • **Human-centered computing** → **Interaction design theory, concepts and paradigms**.

KEYWORDS

social robots, social identity theory, epistemic injustice, trust

ACM Reference Format:
Samantha Stedtler. 2024. Social Injustice, Group Membership and Epistemic Trust in Robots. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/xxxxxx.xxxxxx

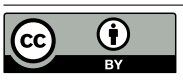
1 INTRODUCTION

Stigmatized groups carry an extra burden, likewise so when it comes to cognitive performance and knowledge production. One of the most known phenomena within social psychology is the 'stereotype threat', where awareness of negative stereotypes of a specific social group lead to its members to underperform in that task [24]. For instance, it has been found that women used to perform more badly in math tasks as a consequence of these stereotypes; even though a supposed male superiority in mathematics has been demonstrated in various studies to not exist [4, 6, 12, 13]. The perception of our identity and worth is intricately connected to how others view us in the social sphere [9, 11, 15]. In addition to this internal threat, there is the experience of being not heard and seen as less competent as a knower, which is called epistemic injustice [7]. Drawing on the example of voice assistants such as Siri or Alexa, a reality where robots give advice while embodying more than just stereotypical white and western identities does not seem too far away. Now imagine a robot giving advice: should it be trusted? It might lack relevant concepts of lived experience, and might lack those more for some groups than for others. As members of oppressed groups, we may be drawn towards other members because we assume that they share similar experiences and perhaps we expect them to function as accomplices or providers of support. Robots however might elicit the same reactions while not offering genuine complicity and embodied knowledge. Thus, I will outline how marginalized groups could be more vulnerable to a specific kind of deception, more specifically, one where identification with the robot because of supposed shared features could lead to unjustified epistemic trust.

2 SOCIAL GROUP DYNAMICS WITH ARTIFICIAL AGENTS

Trusting relationships are steadily woven into the social, material, and embodied dimensions of our world [20]. Maneuvering through these situations means managing an array of encounters, interpretations and expectations, the so-called "social imaginary". The principles that are derived from these influence who is deemed knowledgeable and what can be known.

Social roles and stereotypes seem to be commonly projected onto computers. For instance, participants applied gender stereotypes traditionally associated with human professors to computer tutors, with male tutors being seen as more capable [14]. Gendering AI has been a longstanding practice, illustrated by the historical association of mathematics with male expertise and simultaneously, the recognition of it as the paramount form of intelligence [1]. In addition, voice assistants usually perform tasks associated with feminized domestic labor and are programmed with female voices [18, 22, 25]. Social Identity Theory (SIT) suggests that people organize their lives by grouping themselves into social categories based on their social identity [26]. This group membership seems to be connected to one's self-concept, drive, and behavior through still underexplored psychological and social dynamics [3]. Moreover, it influences sympathy, trust and other positive attitudes towards perceived group members [26]. One example of this is the gender effect observed in Human-Robot Interaction (HRI) as well as



HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03
https://doi.org/10.1145/xxxxxx.xxxxxx

Human-Computer Interaction (HCI), where participants favored robotic or computerized voices that matched with their own gender [5, 14], indicating a higher acceptance of agents that seem to share one's own identity features. So-called similarity-attraction effects have also been identified for personality traits such as introversion vs. extroversion as well [17]. A study by Harwood, Giles, and Ryan [10] underscores the important role of age as a source of social identity. Edwards et al. [3] observed that college students who strongly identified with a higher age group rated an older A.I. voice higher in terms of credibility[1] and social presence than the low age group.

3 EPISTEMIC INJUSTICE AND THE EPISTEMIC ADVANTAGE OF BEING OPPRESSED

Placing epistemic trust in someone entails trusting their competence as a source of knowledge. Related to that, epistemic injustice is a concept that describes how individuals belonging to disadvantaged minorities often encounter challenges in being heard and taken serious as providers of information [7].

Walmsley [27] argued that AI systems pose a risk of perpetuating epistemic injustice by making humans trust those systems more than other humans, since the latter are assigned lower credibility. According to Walmsley, the bias inherent to AI systems could also lead to a form of epistemic injustice where we are hindered from comprehending and expressing our experiences. While I agree that this is a valid risk, I want to focus on another, but related problem. The dominant epistemological and social paradigms in Western contexts emphasize the necessity for knowledge to be objective for its justification [20]. This results in the notion that distrust should be withheld until evidence is presented to motivate such skepticism. Feminist scholarship has criticized this perspective. For instance, Potter [19] asserts that adopting an naïve trusting stance towards others, without exercising discernment, carries risks. There are situations where a cautious attitude of distrust is justified, meaning that marginalized groups could be justified in their distrust towards certain kinds of knowledge and certain sources of knowledge. There is an ongoing debate about whether there is an epistemic advantage in being oppressed. For instance, Dror [2] mentions that there are valid grounds to anticipate that a randomly selected marginalized individual might be in a more favorable epistemic position than a randomly selected non-marginalized person. One example of this is a study which found that on average, 56 percent of Black individuals wrongly assumed that people have equal chances in employment, education, and housing [23]. An even higher average of 81 percent of white individuals held similar false beliefs, highlighting their greater likelihood of subscribing to such misconceptions. It is, as described above, established that we trust people that share our features. One possible explanation in the context of social injustice could be that we trust our group members more because they lived through similar problems and thus have experiential knowledge about being oppressed. It would seem rational then to trust these individuals when they express their knowledge or give us advice. As humans tend to transfer social processes into interactions with robots, this might be unconsciously transferred to a robot, which seems to embody a shared identity, however not in a genuine way, since it is most likely built by people ignorant to this knowledge.

[1]Credibility can be seen as a proxy of epistemic trust.

4 RELATIONAL EFFECTS WITHOUT RELATIONAL RESPONSIBILITIES

Steele, Spencer and Aronson[24] suggested as remedies for the stereotype threat, that friendships, tutors and positive role models can be used to prevent the negative effects of the steretype threat. This can be seen as in line with some efforts of the HRI community to use robots for challenging stereotypes, such as a study by Galatolo et al. [8]. Moreover, conforming to gender norms may not necessarily be the most efficient or advantageous approach: in educational settings, there appeared to be a preference for a mismatch between the gendered characteristics of robots and stereotypical tasks [21]. While some of these ideas certainly work and also do not seem to involve a strong deception or dependence on the robot, studies that investigate the risks of HRI rarely take power imbalences and longterm consequences into account. Steele, Spencer and Aronson [24] call these ways to help minorities and stigmatized groups through friendships, tutors and role models 'relational strategies'. Following that, the question however arises whether a robot can be genuinely relational. It seems like it might be able of having relational effects by providing support and eliciting emotions and trust in users. At the same time, a robot is not capable of being held responsible, and the weight of the relationship and vulnerability of the interaction partner does not matter to its actions. Within moral philosophy, an individual is seen accountable for her actions only if she has control over those actions [16]. This is however not the case for artificial agents, which is why this is commonly understood as a gap in responsibility. Again, this might make oppressed groups more vulnerable to deception since they might ignore their justified reasons for distrust in the institutions and companies building robots as a consequence of the shared identity which the robot seems to embody. Thus, a robot that seems trustworthy based on its group membership and based on being seen as an accomplice might be over-trusted in the knowledge and advice it gives; this in turn could lead to groups which already struggle to have their voices heard and perspectives acknowledged being deceived into taking on advice that has not been made for them.

5 DISCUSSION

In this paper, I argued that we have to consider injustice in social dynamics and knowledge production, such as described in Fricker's concept of epistemic injustice, when implementing identity features into robotic platforms. I focus on the phenomenon of trust where certain forms of distrust can be justified based on experienced injustice, and group membership can be an important tool for transporting knowledge (e.g. in the form of testimony). Social identity features can signal group membership, for instance sharing a gender identity, which can lead to the assumption that one has undergone similar experiences of oppression and can lead to trusting more. Sharing identity features with robots has been shown to lead to similar results, i.e. people liking and trusting the robot more when they identify with them. This can however lead to a new form of deception which oppressed groups might be specifically vulnerable to, where the robot has the same social-relational effects on the person interacting with it, being trusted in its credibility and competence. However, robots do not have the same kind of

responsibilities and knowledge towards these group members: they lack the embodied experience of living through inequality. This leads to the risk of companies, which most likely consist mainly of privileged people, abusing this trust

Based on the literature discussed, I propose three aspects that should be taken into account when designing diverse robot identities:

- Suggestion 1: Users should be able to exert control over the features of a robot, deciding when and in what situations specific features are employed. It is essential for users to recognize that the identity of the robot is not inherently fixed and neither experienced nor "lived" through the robot. It is crucial to emphasize that its features are not hard coded, allowing for flexibility and user customization.
- Suggestion 2: Several, seemingly non-matching identity features should be combined to avoid overtrust and stereotyping. Various identities can be employed for the same task, avoiding, for instance, the exclusive association of female voice assistants solely with domestic labor.
- Suggestion 3: The robot should promote users to rely on their own knowledge to prevent the perpetuation of privileged group's perspectives. Instead of potentially abusing overtrust, the goal should be to foster self-trust.

The first suggestion revolves around the observation in Social Identity Theory (SIT), highlighting that individuals tend to associate themselves based on social identity features. Given that group membership can impact the allocation of trust, and considering the potential for mistakenly assigning trust based on assumed group affiliations, adopting this recommendation could help mitigate risks associated with these SIT-related mechanisms. As mentioned earlier, justified distrust may arise in situations of structural oppression, particularly when directed towards the oppressor and the knowledge derived from their perspective, as noted by Potter[20] . Allowing to switch identity features within the same robot instead of implementing them as permanent properties might make it more apparent that the robot's identity expression might in fact be arbitrary and does not imply specific acquired knowledge through experiences of an identity-coded body. Simultaneously, gaining control over these features can transform an individual's role from a passive user and consumer to an active participant, thus potentially mitigating manifestations of epistemic injustice. This shift is connected to being perceived as a knower, demonstrating competence in modifying the robot's features and understanding what might be necessary and appropriate in a given context. Finally, the third suggestion relates to what Steele, Spencer and Aronson [24] call 'relational strategies'; here, the problem of producing over-trust or an absence of responsibility could be avoided by encouraging the user to trust in their own expertise instead of the robot's.

6 AUTHOR

Samantha Stedtler is a PhD student in Cognitive Science at the LUCS Robotics Group at Lund University. Her focus lies on social robotics, Human-AI Collaboration, 4E cognition and non-dyadic interactions. She investigates expectations, norms, decision-making and dynamics during Human-robot interactions as well as the role of embodiment, situatedness and joint attention. In addition to that, she is also interested in using concepts from feminist technoscience, robophilosophy and AI ethics in her research.

7 ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program - Humanity and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

REFERENCES

[1] Alison Adam. 2006. *Artificial knowing: Gender and the thinking machine.* Routledge.
[2] Lidal Dror. 2023. Is there an epistemic advantage to being oppressed? *Noûs* 57, 3 (2023), 618–640.
[3] Chad Edwards, Autumn Edwards, Brett Stoll, Xialing Lin, and Noelle Massey. 2019. Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions. *Computers in Human Behavior* 90 (2019), 357–362.
[4] Nicole M Else-Quest, Janet Shibley Hyde, and Marcia C Linn. 2010. Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychological bulletin* 136, 1 (2010), 103.
[5] Friederike Eyssel, Dieta Kuchenbrandt, Simon Bobinger, Laura De Ruiter, and Frank Hegel. 2012. 'If you sound like me, you must be more human' on the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction.* 125–126.
[6] Elizabeth Fennema and Julia Sherman. 1977. Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American educational research journal* 14, 1 (1977), 51–71.
[7] Miranda Fricker. 2003. Epistemic justice and a role for virtue in the politics of knowing. *Metaphilosophy* 34, 1-2 (2003), 154–173.
[8] Alessio Galatolo, Gaspar I Melsión, Iolanda Leite, and Katie Winkle. 2023. The right (wo) man for the job? exploring the role of gender when challenging gender stereotypes with a social robot. *International Journal of Social Robotics* 15, 11 (2023), 1933–1947.
[9] Carol Gilligan and ID Voice. 1993. Psychological theory and womens development. *Cambridge, MA* (1993).
[10] Jake Harwood, Howard Giles, and Ellen B Ryan. 1995. Aging, communication, and intergroup theory: Social identity and intergenerational communication. (1995).
[11] Sally Haslanger. 2012. *Resisting reality: Social construction and social critique.* Oxford University Press.
[12] Jane E Hutchison, Ian M Lyons, and Daniel Ansari. 2019. More similar than different: Gender differences in children's basic numerical skills are the exception not the rule. *Child development* 90, 1 (2019), e66–e79.
[13] Alyssa J Kersey, Emily J Braham, Kelsey D Csumitta, Melissa E Libertus, and Jessica F Cantlon. 2018. No intrinsic gender differences in children's earliest numerical abilities. *npj Science of Learning* 3, 1 (2018), 12.
[14] Eun Ju Lee, Clifford Nass, and Scott Brave. 2000. Can computer-generated speech have gender? An experimental test of gender stereotype. In *CHI'00 extended abstracts on Human factors in computing systems.* 289–290.
[15] Federica Liveriero et al. 2019. The social bases of self-respect. Political equality and epistemic injustice. *Phenomenology and Mind* 16 (2019), 90–101.
[16] Michael McKenna. 2008. Putting the lie on the control condition for moral responsibility. *Philosophical Studies* 139 (2008), 29–37.
[17] Clifford Ivar Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship.* MIT press Cambridge.
[18] Giulia Perugia and Dominika Lisy. 2023. Robot's gendering trouble: a scoping review of gendering humanoid robots and its effects on HRI. *International Journal of Social Robotics* (2023), 1–29.
[19] Nancy Nyquist Potter. 2002. *How can I be trusted?: a virtue theory of trustworthiness.* Rowman & Littlefield.
[20] Nancy Nyquist Potter. 2020. Interpersonal trust. In *The routledge handbook of trust and philosophy.* Routledge New York, 243–255.
[21] Natalia Reich-Stiebert and Friederike Eyssel. 2017. (Ir) relevance of gender? On the influence of gender stereotypes on learning with a robot. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction.* 166–176.
[22] Jennifer Rhee. 2018. *The robotic imaginary: The human and the price of dehumanized labor.* U of Minnesota Press.
[23] Jim Sidanius and Felicia Pratto. 2001. *Social dominance: An intergroup theory of social hierarchy and oppression.* Cambridge University Press.
[24] Claude M Steele, Steven J Spencer, and Joshua Aronson. 2002. Contending with group image: The psychology of stereotype and social identity threat. In *Advances in experimental social psychology.* Vol. 34. Elsevier, 379–440.

[25] Yolande Strengers and Jenny Kennedy. 2021. *The smart wife: Why Siri, Alexa, and other smart home devices need a feminist reboot.* Mit Press.
[26] Henri Tajfel. 1978. Social categorization, social identity and social comparison. *Differentiation between social group* (1978), 61–76.
[27] Joel Walmsley. 2023. "Computer Says No": Artificial Intelligence, Gender Bias, and Epistemic Injustice. In *Feminist Philosophy and Emerging Technologies*. Routledge, 249–263.