Using Robot Social Agency Theory to Understand Robots' Linguistic Anthropomorphism

Cloe Z. Emnett
cemnett@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Terran Mott
terranmott@mines.edu
Colorado School of Mines
Golden, Colorado, USA

Tom Williams
twilliams@mines.edu
Colorado School of Mines
Golden, Colorado, USA

ABSTRACT

Robots' use of natural language is one of the key factors that leads humans to anthropomorphize them. But it is not yet well understood what types of language most lead to such language-based anthropomorphization (or, *Linguistic Anthropomorphism*). In this paper, we present a brief literature survey that suggests six broad categories of linguistic factors that lead humans to anthropomorphize robots: autonomy, adaptability, directness, politeness, proportionality, and humor. By contextualizing these six factors through the lens of Jackson and Williams' Theory of Social Agency for Human-Robot Interaction, we are able to show how and why these particular factors are those responsible for language-based robot anthropomorphism.

CCS CONCEPTS

• **Computer systems organization → Robotics**; • **Human-centered computing → Human computer interaction (HCI)**; *HCI theory, concepts and models.*

KEYWORDS

linguistic anthropomorphism, social agency

ACM Reference Format:
Cloe Z. Emnett, Terran Mott, and Tom Williams. 2024. Using Robot Social Agency Theory to Understand Robots' Linguistic Anthropomorphism. 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, 6 pages. https://doi.org/10.1145/3610978.3640747

1 MOTIVATION

In the future, social robots will likely continue to build the capability to exist alongside humans in familiar environments. Robots can provide physical and social assistance in healthcare [27, 39], education [21, 48], therapy [11], and in the home [15]. But for social robots to be effective in these unpredictable and unconstrained settings, they must competently respond to a variety of complex, potentially high-stakes interactions. For example, robots will inevitably confront ethically sensitive interactions, such as when they receive unethical commands [22] or witness bias [37, 51]. Across these different contexts, the way that robots signal human-likeness can impact whether robots are perceived as appropriate [37] and trustworthy [19, 45, 53]. Therefore, roboticists must carefully consider the design cues they implement that emphasize a robot's human-likeness.

One of the key ways that robots signal human-likeness is by using human-like language cues. This *linguistic anthropomorphism* can have a number of key downstream effects on interaction. Robots that mimic human linguistic patterns can promote encouraging [17] pro-social [31] interactions. Robots are more successful and acceptable collaborators when they have human-like social competence and are sensitive to human sociocultural norms [37, 40] and social roles [47]. Robots that utilize human-like linguistic strategies can successfully, yet tactfully reject unethical commands [25**?**], address bias [37, 51], and handle high stakes interactions in trustworthy ways [19, 45, 53]. However, robot's use of human-like language also presents potential drawbacks. In particular, robot's use of human-like language can sometimes conflict with humans' assessment of robot's social standing. For example, it can be inappropriate for robots to use human-like linguistic cues when doing so might be uncanny [8, 10] or an inappropriate role for robots to take on [33, 37]. Similarly, robots may be perceived as uncanny or untrustworthy if they misuse human-like language features in particular contexts, such as using indirect speech when giving critical directions for driving [52].

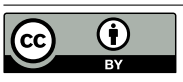
While it is broadly recognized that there are many dimensions of robot language use that leads interactants to anthropomorphize them, the field lacks a coherent theory of what those factors are, and why they lead to anthropomorphization. In this review, we thus investigate the research question: ***What characteristics of linguistic anthropomorphism are relevant human-robot interaction?*** To answer this question, we survey and organize previous work that has examined the characteristics of natural language that make robots be perceived as more human-like.

A surprising outcome of this survey is a clear mapping from the antecedents of linguistic anthropomorphism explored in the literature, and the key dimensions of social agency proposed by Jackson and Williams [24] in their *Theory of Social Agency for Human-Robot Interaction.*

1.1 Jackson and Williams' Theory of Social Agency for Human-Robot Interaction

Due to the importance of Jackson and Williams' framework for the organization and interpretation of our results, we will briefly summarize it before describing the results of our survey.

When humans interact with robots in the wild, especially in these high stakes situations, they must make decisions about the extent to which robots are social and moral others [46]. As one dimension of



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

ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3640747

this social categorization, people (explicitly or implicitly) categorize others as *agents* based on a number of key observable factors.

Under Floridi and Sanders [14]'s general theory of artificial agency, three key features contribute to whether an entity is an agent from the perspective of a particular observer: whether it is interactive (able to act on its environment and be acted on in return), autonomous (able to make its own decisions), and adaptable (able to learn over time). Jackson and Williams [24]'s *Theory of Social Agency for Human-Robot Interaction* builds on this framework to suggest that for robots to be *social* agents, they must (1) be agents according to this definition, and (2) have the clear capability for *social action*, which they define as the ability to threaten or affirm the Face (or social standing) of others [5].

As we will see, this framework provides a clear explanation for what factors lead to robot's linguistic anthropomorphism, and in turn, provides new evidence for Jackson and Williams [24]'s theory. As such, by drawing this connection, our survey lays the foundation for future experimental work investigating the effect of linguistic anthropomorphism on a robots trustworthiness, credibility, and social competence in ethically sensitive interactions, by providing a clear framework through which linguistic anthropomorphism can be manipulated and assessed.

2 METHODS

The goal of our review was to investigate key characteristics of linguistic anthropomorphism for social robots. We searched for "anthropomorphism" and "linguistic anthropomorphism" through the ACM and IEEE digital libraries, and evaluated resulting papers for their fit with our topic. We particularly gathered papers studying human-like language in sensitive interactions, including command rejection, providing health information, moderating conflict, and potentially dangerous situations like driving and security. This process resulted in 36 papers from HRI, HCI, AutomotiveUI, RO-MAN, and Conversational Agents. This included 23 papers from the past 5 years and 13 that were more than 5 years old.

3 A REVIEW OF LINGUISTIC ANTHROPOMORPHISM IN HRI

In this section, we review key characteristics of linguistic anthropomorphism through the lens of social agency theory. We describe how each characteristic emphasizes human-likeness in natural language interaction and note its potential advantages and risks. Furthermore, we note how each characteristic may contribute to a user's assessment of a robot's agency or capacity for social action.

3.1 Adaptability through Personalization

In our review of literature on factors that might serve as the antecedents of linguistic anthropomorphism, we found a number of papers that suggest the importance of a robot's perceived adaptability during conversations, such as robots' ability to integrate personal details about other interlocutors into discourse [45]. Conversational adaptability can be demonstrated in a variety of ways such as adapting to the initial situation, changes in the situation, the person communicating with,etc.. Often, this comes in the form of *personalization*.

When a robot adapts to information acquired during conversation with the relevant human interlocutor, there is a corresponding increase in perceived anthropomorphism. While adaptability does not necessarily result in elevated trust levels, a higher degree of anthropomorphism tends to correlate with heightened trust [45].

The longitudinal use of personalization has demonstrated its capacity to enhance cooperation, rapport, and engagement [30]. Even the simple use of lexical entrainment [3], in which the phrases and speaking patterns of a human conversational partner are mirrored by a robot, may lead to increased anthropomorphism and positive downstream effects.

However, when a robot engages in personalization based on its own experiences, whether authentic or fabricated, lower levels of likability are sometimes observed. This approach may imply the robot's aspiration to establish social equality with humans, an interpretation negatively received due to its deviation from the robot's purported original intention—prioritizing human needs [29]. Although this manifestation of adaptability enhances anthropomorphism, it concurrently diminishes the likelihood of compliance from the human participant, fostering negative perceptions of the robot.

Critically, these cases of adaptability through personalization may simultaneously lead to anthropomorphism both because they signal agency, and because they signal capacity for social action. Adaptability is a key facet of agency under Floridi and Sanders [14]'s theory of artificial agency. Moreover, adaptability *in the form of personalization* may also be perceived as social action because a robot's recall of its interaction partner's personal details may affirm their Face, or social standing, by emphasizing familiarity [5].

3.2 Autonomy through Assertiveness

Next, we found a number of papers that suggest the importance of a robot's perceived assertiveness, or confidence in its decisions, for robot anthropomorphization. In particular, assertiveness appears to correlate positively with anthropomorphism, in a way that engenders greater trust [19]. For example, when choosing a voice for a self-driving car, assertive, human-like voices can garner more attention from drivers than less-anthropomorphic machine-like voices[52]. Similarly, in high-stakes job interviews facilitated by a robot interviewer, assertiveness can add anthropomorphism to the design robot interviewer's personality, correlating with heightened engagement and attentiveness [53].

When a conversational agent (CA) exhibits heightened confidence in conversation, it is perceived with increased levels of trustworthiness [38]. Moreover, assertiveness can be strategically employed to convey a sense of authority. Instances demanding high cognitive engagement from the human demonstrate improved performance when conversational styles embody increased authority. This heightened performance underscores a greater level of trust and likability [34]. Importantly, higher levels of authority yield heightened trust [19, 34].

On the other hand, assertiveness can also lead to decreased likability when it is construed as aggressive [1]. Robots' use of assertiveness must thus be combined with mutual respect to mitigate the potential perception of assertiveness as aggressiveness.

Component	Characteristic	Definition
Agency *and* Social Action	Adaptability (through Personalization)	Referencing past experiences relating to the robots fictional past or past interactions with humans [3, 4, 29, 30, 45]
	Autonomy (through Assertiveness)	Portraying self-assurance [19, 34, 38, 52, 53]
Social Action	Directness	Using, or not using, hedge or discourse markers to influence the level of inference needed to interpret a request [35, 41, 44]
	Politeness	Including words that make a statement more respectful and considerate of other [9, 18, 20, 40]
	Proportionality	Utilizing the above strategies in a way that is reflective of the situation the interaction takes place [2, 12, 13, 22]
Overtly Human Behavior	Humor	Incorporating causal references to people or pop culture [28, 42]

Table 1: Antecedents of Linguistic Anthropomorphism

Increased perceived trustworthiness also has it's risks in situations where the robot is not worthy of the perceived trustworthiness.

Critically, this assertiveness may convey the speaker's level of *autonomy*, a key factor in determining a robot's agency [14]; and, assertiveness may convey the speaker's potential for social action, if it is construed as Face-threatening and aggressive [5]. As such, these cases of autonomy through assertiveness may simultaneously lead to anthropomorphism both because they signal agency, and because they signal capacity for social action.

3.3 Directness

In direct speech, the intended meaning of speech acts corresponds with their logical meaning. In contrast, humans use indirect speech acts to blur their intended meaning for some reason, such as to be polite. Robots use of indirect language can correspond to a heightened level of anthropomorphism [41].

Opting for indirect speech can yield positive effects on the perceived qualities of a robot. Notably, incorporating hedge and discourse markers into word choice can enhance a robot's image, making it appear more considerate and likable [44]. Hedge and discourse markers serve to temper statements, imparting a more casual tone to the communication. Hedge markers, representing a form of negative politeness, will be expounded upon in the subsequent section. In scenarios involving requests, the employment of indirect speech has been associated with increased compliance and elevated perceptions of trustworthiness [41].

Conversely, implicit speech may assume a negative connotation in high-stakes situations [35]. For instance, in the context of driving instructions, direct speech is favored for its perceived utility. In situations fraught with elevated risk, where compliance with the robot's instructions carries significant consequences, explicit speech may prove more effective despite the general preference for indirect speech.

If implemented in ways that are appropriate for an interaction context, robot's use of indirect speech may contribute to users' perception of their capacity for social action. Among humans, indirect speech is an important linguistic tool for minimizing face threats and attending to the social standing of others, such as when softening harsh statements [5].

3.4 Politeness

Politeness in human language involved a variety of different linguistic cues, ranging from pragmatic strategies (such as gratitude, deference, or appeals in-group membership) to syntactic choices (such as plural pronouns and passive voice) [9]. Given the deeply ingrained human nature of politeness, heightened levels of this trait align with increased anthropomorphism. The expectation for robots to adhere to human social conventions further underscores the relevance of politeness in robotic communication [40].

Consequently, employing politeness strategies in robotic interactions may contribute to a more positive perception of robots [18, 20]. Existing evidence suggests that heightened politeness in robots fosters more constructive interactions, although its impact on the acceptance of a robot's non-compliance remains an open question. Employing politeness to temper a statement enhances its perceived receptivity [20].

However, the use of politeness to create human-like robot speech can have potential drawbacks. People expect to have more social power over robots than they do over humans in equivalent roles [33], which is a main determinant of politeness norms [9, 32]. Therefore, robots that mimic human-like politeness may be perceived as disingenuous. It can be inappropriate for robots to use linguistic cues which allude to inherently human experiences or characteristics [7, 43]. Robots can be perceived as uncanny when they use human linguistic politeness in ways that users' feel is inappropriate for non-human entities [8, 10, 49]. For example, it may be deceitful for a robot to be polite by referencing emotions it cannot have [6].

Politeness is an essential component of a robot's perceived capacity for social action because it is used to minimize possible face threats [5, 24]. When a robot is utilizing a politeness strategy, it is ensuring that the message coming across is respectful and does not negatively impact what the human conversational partner thinks of the robot. Politeness represents a communication tool commonly employed by humans in interpersonal interactions. Its significance becomes particularly pronounced in high-stakes exchanges, where a statement lacking in politeness might be construed as critical, harsh, or even hostile.

3.5 Proportionality

Proportionality is a linguistic behavior that humans use to tune the severity of their language in order to ensure that the severity of their response corresponds to the severity of the situation at hand. For example, humans often select proportional responses when rebuking others or refusing a request [16, 26]. Proportionality is a key component of robots' human-like social competence in sensitive situations [37], such as addressing inappropriate actions [23, 51]. Because proportionality is a linguistic strategy for modulating the harshness of speech, it is closely related to linguistic politeness and directness [5].

When a robot can respond proportionally, it demonstrates an ability to navigate social norms, mirroring human behavior. This adherence to established norms is crucial, as it correlates with heightened levels of trustworthiness [12], acceptability [13], and credibility [2] in robotic interactions. An improved perception along these dimensions increases the likelihood of humans accepting a robot's non-compliance. Proportionality can enable robots to tactfully reject unethical commands [23, 25] and address bias [37, 50, 51]. A failure to align non-compliance responses with proportionality may lead to robot being perceived as over or under-severe [23, 37].

Proportionality is a key component of robots' human-like social competence in sensitive situations [37]. The ability to modulate the harshness of speech makes it closely related to linguistic politeness and directness [5]. In this way, a robot's ability to be proportional contributes to its capability for social action by attending to the Face of others.

3.6 Humor

Humor in language is intended to illicit amusement from others. Higher levels of perceived humor in robot interactions tend to correlate with higher levels of anthropomorphism. Robots can use humor to facilitate ice-breakers in conversation [42] and to create a more casual environment by using causal terms to reference others, for example "dude" [28]. The exploration of humor in high-stakes situations merits further investigation, as elevated levels of humor hold promise for conflict resolution.

Humor, as an interaction design tool in HRI, demonstrates efficacy in alleviating tension [28]. It proves particularly valuable in mitigating the discomfort a human might experience when facing denial of a request and can improve the level of perceived likability [42]. Because robots can use humor to create a more casual environment, it should be used with caution. To make light of serious situation might not be perceived positively. Judicious implementation of humor is essential and should be contingent upon the gravity of the command to which the robot is non-compliant [28]. Notably, humor may not always be suitable when the command in question warrants seriousness. In such cases, its application may be more aptly reserved for post-interaction moments, serving as a means to repair rapport.

In some case, humor can influence social dynamics by affirming or threatening individuals involved in the interaction [5]. Yet even these types of humor can backfire due to being perceived as too overtly human-like [36]. This suggests that humor may be better categorized as an overtly human form of social action than being fit into the social action framework used to reason about non-human agents.

4 DISCUSSION

Analyzing these six factors through the lens of Jackson and Williams' Theory of Social Agency for Human-Robot Interaction enables us to fully explore why these factors have strong relationships to anthropomorphism. Adaptability through personalization and autonomy through assertiveness both have the capability of portraying social agency and action. Directness, politeness, proportionality are all examples of a robots ability for social action. Our results have significant implications for the design of social robots, given the observed effects of these dimensions of social agency on key human factors such as perceived trustworthiness, likeability, and social competence.

More generally, though, our analysis shows that *the reason* why different types of verbal behaviors result in linguistic anthropomorphism may be because those behaviors demonstrate key dimensions of robot social agency. Future work is needed to concretely test this theory by directly testing the extent to which these dimensions of social agency mediate the ways these strategies lead to perceived human-likeness.

Moreover, the fact that certain strategies, like humor, may backfire due to being perceived as "too humanlike", suggests that future research might examine (1) whether *overuse* of any of these strategies might lead to uncanny valley effects[8], (2) which specific types of humor might increase the perception of specific aspects of social agency, and (3) which specific types of humor might be perceived as "too humanlike".

5 CONCLUSION

This mini-review investigates linguistic anthropomorphism, presenting a framework that explores its impact on robots' trustworthiness, credibility, and acceptance. We consider various facets of linguistic anthropomorphism, including assertiveness, adaptability, humor, directness, politeness, and proportionality. Contextualizing these six factors through the lens of Jackson and Williams' Theory of Social Agency for Human-Robot Interaction, we show how these particular factors may influence assessments of robots' social agency. This work establishes a framework for future experimental inquiries on the effects of linguistic anthropomorphism in human-robot interactions, particularly when navigating non-compliance or high stakes scenarios.

ACKNOWLEDGMENTS

This work was funded in part by Young Investigator award FA9550-20-1-0089 from the United States Air Force Office of Scientific Research.

REFERENCES

[1] Siddharth Agrawal and Mary-Anne Williams. 2018. Would You Obey an Aggressive Robot: A Human-Robot Interaction Field Study. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (Nanjing, China). IEEE Press, 240–246. https://doi.org/10.1109/ROMAN.2018.8525615

[2] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. 2015. Effects of Culture on the Credibility of Robot Speech: A Comparison between English and Arabic. In *Proceedings of the Tenth Annual ACM/IEEE*

International Conference on Human-Robot Interaction (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/2696454.2696464

[3] Deepali Aneja, Rens Hoegen, Daniel McDuff, and Mary Czerwinski. 2021. Understanding Conversational and Expressive Style in a Multimodal Embodied Conversational Agent. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 102, 10 pages. https://doi.org/10.1145/3411764.3445708

[4] Susan E. Brennan and Justina O. Ohaeri. 1994. Effects of Message Style on Users' Attributions toward Agents. In *Conference Companion on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 281–282. https://doi.org/10.1145/259963.260492

[5] Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

[6] Herbert Clark and Kerstin Fischer. 2022. Social robots as depictions of social agents - Behavioral and Brain Sciences (forthcoming). *Behavioral and Brain Sciences* 2022 (07 2022), 1–33.

[7] Leigh Clark. 2018. Social Boundaries of Appropriate Speech in HCI: A Politeness Perspective. In *Proceedings of British HCI*.

[8] Leigh Clark, Abdulmalik Yusuf Ofemile, and Benjamin Cowan. 2020. *Exploring Verbal Uncanny Valley Effects with Vague Language in Computer Speech*. 317–330. https://doi.org/10.1007/978-981-15-6627-1_17

[9] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Daniel Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Annual Meeting of the Association for Computational Linguistics*.

[10] Autumn Edwards, Chad Edwards, and Andrew Gambino. 2020. The Social Pragmatics of Communication with Social Robots: Effects of Robot Message Design Logic in a Regulative Context. *International Journal of Social Robotics* 12 (08 2020).

[11] Saad Elbeleidy, Terran Mott, Dan Liu, and Tom Williams. 2022. Practical Considerations for Deploying Robot Teleoperation in Therapy and Telehealth. In *Int'l Symposium on Robot-Human Interactive Communication*.

[12] Vanessa Evers, Heidy C. Maldonado, Talia L. Brodecki, and Pamela J. Hinds. 2008. Relational vs. Group Self-Construal: Untangling the Role of National Culture in HRI. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction* (Amsterdam, The Netherlands) *(HRI '08)*. Association for Computing Machinery, New York, NY, USA, 255–262. https://doi.org/10.1145/1349822.1349856

[13] Imran Fanaswala, Brett Browning, and Majd Sakr. 2011. Interactional Disparities in English and Arabic Native Speakers with a Bi-Lingual Robot Receptionist. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (Lausanne, Switzerland) *(HRI '11)*. Association for Computing Machinery, New York, NY, USA, 133–134. https://doi.org/10.1145/1957656.1957697

[14] Luciano Floridi and Jeff W Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14 (2004), 349–379.

[15] Jodi Forlizzi. 2007. How Robotic Products Become Social Products: An Ethnographic Study of Cleaning in the Home. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (Arlington, Virginia, USA) *(HRI '07)*. Association for Computing Machinery, 129–136.

[16] Erving Goffman. 1967. *Interaction Ritual: Essays in Face-to-Face Behavior*.

[17] Stephan Hammer, Birgit Lugrin, Sergey Bogomolov, Kathrin Janowski, and Elisabeth André. 2016. Investigating Politeness Strategies and Their Persuasiveness for a Robotic Elderly Assistant. In *Proceedings of the 11th International Conference on Persuasive Technology - Volume 9638* (Salzburg, Austria) *(PERSUASIVE 2016)*. Springer-Verlag, 315–326.

[18] Stephan Hammer, Birgit Lugrin, Sergey Bogomolov, Kathrin Janowski, and Elisabeth André. 2016. Investigating Politeness Strategies and Their Persuasiveness for a Robotic Elderly Assistant. In *Proceedings of the 11th International Conference on Persuasive Technology - Volume 9638* (Salzburg, Austria) *(PERSUASIVE 2016)*. Springer-Verlag, Berlin, Heidelberg, 315–326. https://doi.org/10.1007/978-3-319-31510-2_27

[19] Christina N. Harrington and Lisa Egede. 2023. Trust, Comfort and Relatability: Understanding Black Older Adults' Perceptions of Chatbot Design for Health Information Seeking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (, Hamburg, Germany,) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 120, 18 pages. https://doi.org/10.1145/3544548.3580719

[20] Thomas Holtgraves. 2021. Understanding Miscommunication: Speech Act Recognition in Digital Contexts. *Cognitive science* 45 (10 2021), e13023. https://doi.org/10.1111/cogs.13023

[21] Deanna Hood, Séverin Lemaignan, and Pierre Dillenbourg. 2015. When Children Teach a Robot to Write: An Autonomous Teachable Humanoid Which Uses Simulated Handwriting. *ACM/IEEE International Conference on Human-Robot Interaction* 2015 (03 2015), 83–90. https://doi.org/10.1145/2696454.2696479

[22] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. 2019. Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 499–505. https://doi.org/10.1145/3306618.3314241

[23] Ryan Blake Jackson, Ruchen Wen, and Tom Williams. 2019. Tact in Noncompliance: The Need for Pragmatically Apt Responses to Unethical Commands. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*.

[24] Ryan Blake Jackson and Tom Williams. 2021. A Theory of Social Agency for Human-Robot Interaction. *Frontiers in Robotics and AI* (2021).

[25] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the Role of Gender in Perceptions of Robotic Noncompliance. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[26] Danette Ifert Johnson, Michael E. Roloff, and Melissa A. Riffee. 2004. Politeness theory and refusals of requests: Face threat as a function of expressed obstacles. *Communication Studies* (2004).

[27] Khari Johnson. 2022. Hospital Robots Are Helping Combat a Wave of Nurse Burnout.

[28] Malte F. Jung, Nikolas Martelaro, and Pamela J. Hinds. 2015. Using Robots to Moderate Team Conflict: The Case of Repairing Violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 229–236. https://doi.org/10.1145/2696454.2696460

[29] Hyeji Kim, Inchan Jung, and Youn-kyung Lim. 2022. Understanding the Negative Aspects of User Experience in Human-Likeness of Voice-Based Conversational Agents. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (<conf-loc>, <city>Virtual Event</city>, <country>Australia</country>, </conf-loc>) *(DIS '22)*. Association for Computing Machinery, New York, NY, USA, 1418–1427. https://doi.org/10.1145/3532106.3533528

[30] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis, and Sarun Savetsila. 2012. Personalization in HRI: A Longitudinal Field Experiment. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (Boston, Massachusetts, USA) *(HRI '12)*. Association for Computing Machinery, New York, NY, USA, 319–326. https://doi.org/10.1145/2157689.2157804

[31] Namyeon Lee, Jeonghun Kim, Eunji Kim, and Ohbyung Kwon. 2017. The Influence of Politeness Behavior on User Compliance with Social Robots in a Healthcare Service Setting. *International Journal of Social Robotics* 9 (11 2017).

[32] Geoffrey Leech. 2014. The Pragmatics of Politeness. *The Pragmatics of Politeness* (07 2014), 1–368.

[33] Eleonore Lumer and Hendrik Buschmeier. 2022. Perception of Power and Distance in Human-Human and Human-Robot Role-Based Relations. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[34] Gianpaolo Maggi, Elena Dell'Aquila, Ilenia Cucciniello, and Silvia Rossi. 2021. "Don't Get Distracted!": The Role of Social Robots' Interaction Style on Users' Cognitive Performance, Acceptance, and Non-Compliant Behavior. *International Journal of Social Robotics* 13 (12 2021). https://doi.org/10.1007/s12369-020-00702-4

[35] Tomoki Miyamoto, Daisuke Katagami, Takahiro Tanaka, Hitoshi Kanamori, Yuki Yoshihara, and Kazuhiro Fujikake. 2021. Should a Driving Support Agent Provide Explicit Instructions to the User? Video-Based Study Focused on Politeness Strategies. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (Virtual Event, Japan) *(HAI '21)*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/3472307.3484160

[36] Terran Mott, Aaron Fanganello, and Tom Williams. 2024. What a Thing to Say! Which Linguistic Politeness Strategies Should Robots Use in Noncompliance Interactions?. In *ACM/IEEE International Conference on Human-Robot Interaction*.

[37] Terran Mott and Tom Williams. 2023. Confrontation and Cultivation: Understanding Perspectives on Robot Responses to Norm Violations. In *Proceedings of the IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*.

[38] Marissa Radensky, Julie Anne Séguin, Jang Soo Lim, Kristen Olson, and Robert Geiger. 2023. "I Think You Might Like This": Exploring Effects of Confidence Signal Patterns on Trust in and Reliance on Conversational Recommender Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 792–804. https://doi.org/10.1145/3593013.3594043

[39] J Saldien, K Goris, B Vanderborght, B Verrelst, R Van Ham, and D Lefeber. 2006. ANTY : The development of an intelligent huggable robot for hospitalized children. (2006), 6.

[40] Maha Salem, Micheline Ziadee, and Majd Sakr. 2014. Marhaba, How May i Help You? Effects of Politeness and Culture on Robot Acceptance and Anthropomorphization. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (Bielefeld, Germany) *(HRI '14)*. Association for Computing Machinery, New York, NY, USA, 74–81. https://doi.org/10.1145/2559636.2559683

[41] Shane Saunderson and Goldie Nejat. 2021. Robots Asking for Favors: The Effects of Directness and Familiarity on Persuasive HRI. *IEEE Robotics and Automation Letters* 6, 2 (2021), 1793–1800. https://doi.org/10.1109/LRA.2021.3060369

[42] Moonyoung Tae and Joonhwan Lee. 2020. The Effect of Robot's Ice-Breaking Humor on Likeability and Future Contact Intentions. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 462–464. https://doi.org/10.1145/3371382.3378267

[43] Marina Terkourafi. 2005. Beyond the Micro-level in Politeness Research. *Journal of Politeness Research-language Behaviour Culture* 1 (07 2005), 237–262.

[44] Cristen Torrey, Susan Fussell, and Sara Kiesler. 2013. How a Robot Should Give Advice. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction* (Tokyo, Japan) *(HRI '13)*. IEEE Press, 275–282.

[45] Rebecca Wald, Evelien Heijselaar, and Tibor Bosse. 2021. Make Your Own: The Potential of Chatbot Customization for the Development of User Trust. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 382–387. https://doi.org/10.1145/3450614.3463600

[46] Kara Weisman. 2022. Extraordinary entities: Insights into folk ontology from studies of lay people's beliefs about robots.. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*.

[47] Ruchen Wen, Zhao Han, and Tom Williams. 2022. Teacher, Teammate, Subordinate, Friend: Generating Norm Violation Responses Grounded in Role-Based Relational Norms. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[48] Jacqueline M. Kory Westlund, Hae Won Park, Randi Williams, and Cynthia Breazeal. 2018. Measuring Young Children's Long-Term Relationships with Social Robots. In *Proceedings of the 17th ACM Conference on Interaction Design and Children* (Trondheim, Norway) *(IDC '18)*. Association for Computing Machinery, New York, NY, USA, 207–218. https://doi.org/10.1145/3202185.3202732

[49] Tom Williams, Priscilla Briggs, and Matthias Scheutz. 2015. Covert Robot-Robot Communication: Human Perceptions and Implications for Human-Robot Interaction. *J. Hum.-Robot Interact.* (sep 2015), 24–49.

[50] Katie Winkle, Ryan Blake Jackson, Gaspar Isaac Melsión, Dražen Brščić, Iolanda Leite, and Tom Williams. 2022. Norm-Breaking Responses to Sexist Abuse: A Cross-Cultural Human Robot Interaction Study. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[51] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*.

[52] Priscilla N. Y. Wong, Duncan P. Brumby, Harsha Vardhan Ramesh Babu, and Kota Kobayashi. 2019. Voices in Self-Driving Cars Should Be Assertive to More Quickly Grab a Distracted Driver's Attention. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Utrecht, Netherlands) *(AutomotiveUI '19)*. Association for Computing Machinery, New York, NY, USA, 165–176. https://doi.org/10.1145/3342197.3344535

[53] Michelle X. Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting Virtual Agents: The Effect of Personality. *ACM Trans. Interact. Intell. Syst.* 9, 2–3, Article 10 (mar 2019), 36 pages. https://doi.org/10.1145/3232077