Purposeful Failures as a Form of Culturally-Appropriate Intelligent Disobedience during Human-Robot Social Interaction

Casey C. Bennett
School of Intelligence Computing
Hanyang University
Seoul, Korea
cabennet@hanyang.ac.kr

Benjamin Weiss
Quality and Usability Lab
Technische Universität
Berlin, Germany
benjamin.weiss@tu-berlin.de

ABSTRACT

Human-robot interaction (HRI) can suffer from *breakdowns* that are often regarded as "failures" by roboticists. Here, however, we argue that such breakdowns can be sometimes perceived as a type of *defiance* that signals more socially intelligent behavior rather than less, depending on the culture and linguistic environment within which they occur. We present recent research evidence supporting this viewpoint, based on HRI experiments comparing English speakers and Korean speakers. Counterintuitively, occasional culturally-appropriate forms of disobedience may in fact be a desirable design feature for social robots in the future.

KEYWORDS

human robot interaction, failures, games, social interaction, inhibition of return, autonomous agents

ACM Reference Format:
Casey C. Bennett and Benjamin Weiss. 2022. Purposeful Failures as a Form of Culturally-Appropriate Intelligent Disobedience during Human-Robot Social Interaction. In *RAD-AI Workshop of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Auckland, New Zealand, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

An idea gaining broader attention recently is the concept of social interaction breakdown during human-robot interaction (HRI) studies. Within the HRI field, these breakdowns are often regarded as "failures" [2, 17]. While on the surface that may be true in principle, there is a growing body of HRI research that focuses on cross-cultural robotics, with the aim of conducting the exact same experiment in multiple geographic locations as well as multiple languages/cultures [3, 29, 30]. This involves both physically embodied robots as well as virtual avatars. Results from that research have led us to an interesting notion, that **some breakdowns may not be regarded as failures at all in certain cultural environments, but rather may be perceived as more intelligent behavior in the form of occasional defiance of existing social norms**. Moreover, what may be viewed as "normal" behavior in one cultural environment may be seen as "defiant" in another [8]. In that sense, such *perceived disobedience* may in fact create an opportunity to design more effective social robots.

RAD-AI Workshop of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Auckland, New Zealand.

A good example of this is a comparison between East Asian and Western cultures, for which there is significant existing research [20]. In particular, what constitutes appropriate robot/agent behavior in those environments is defined by several dimensions: high-context vs. low-context cultures, communal vs. individualistic cultures, Confucian power hierarchies vs. Western power hierarchies, etc. [4, 23, 29]. A broad review of such existing research can be found in Lim et al. 2021 [24].

Such cultural differences are also embedded into our spoken languages, otherwise known as *linguistic relativity* [12, 14]. The basic argument is that the language we speak affects our habitual behaviors and worldview [7], much like a lens warps visual perception. In bilingual speakers, this can constitute cognitive differences depending on the language they are actively speaking in [21]. This lens metaphor extends to expectations, e.g. the right amount of gaze or verbal backchannel depends on (culturally different) roles and hierarchies [19]. If associated behavior does not meet the expectation, it can irritate a dialogue partner so much that it requires explicit clarification, disrupting the flow of communication [13]. This, however, is culturally dependent, e.g. a strong indicator of a breakdown in some Western cultures, silence, can represent a very appropriate contribution to dialogue in high-context cultures [9].

Below we present a brief summary of some recently completed research related to these topics and this workshop on rebellion and disobedience in AI [11, 25]. We then discuss our perspectives on how research around cross-cultural robotics can contribute to the broader discussion of intelligent disobedience and its applications within the HRI field.

2 METHODS

Our own research lends further evidence to this position. In particular, one of our recently concluded studies focuses on the concept of *Social Inhibition of Return* (social IOR), which is based on IOR models from various human sensory functions such as vision [26]. The basic idea here is that there are mechanisms in the brains of naturally intelligent organisms (including humans) that inhibit us from repeating the same behavior in a short period of time (e.g. 2-3 seconds) in order to maximize task efficiency (e.g. during visual "information foraging") [18]. A failure in these mechanisms is thought to play a role in human mental illness, such as obsessive-compulsive disorder. These mechanisms are also important to produce fluid natural behavior, rather than repetitive "robot-like behavior" in humans [26]. In an HRI context, the removal of social IOR causes the agent to engage in repetitive speech behaviors as well as increase the chance of "talking over" the human interactor. **Such behavior theoretically could be seen as *defying* typical social norms,**

or perhaps even "aggressive" social behavior by the agent, rather than trying to engage the human on their terms (as is the typical approach in HRI). We posit that such defiance may affect perceived disobedience of the robot by the human, as defined in Section 1.

In the recent study, we experimented with social IOR, by having a virtual avatar play a social survival video game called "Don't Starve Together" with a human player. The agent was capable of autonomous speech interactions during gameplay (i.e. *Social AI*), which was developed specifically for cooperative game paradigms during previous studies [5]. This specific Social AI was capable of hundreds of different speech utterances covering 46 different utterance categories, each related to a particular game situation (e.g. collecting resources, fighting monsters, deciding where to go next) organized as a hierarchy with several levels. Those speech utterances were both self-generated based on internal logic of the Social AI, as well as responses to human player speech via automatic speech recognition (ASR). The speech responses were similar in both English and Korean (i.e. the robot was bilingual, in essence).

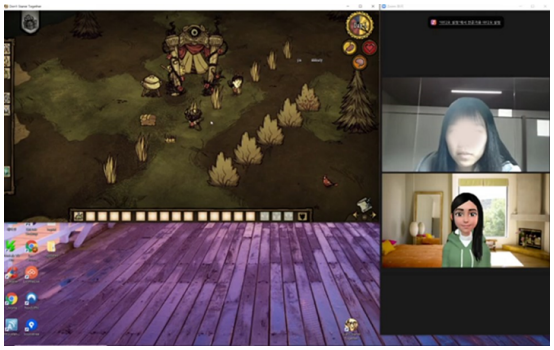

Figure 1: Gameplay example during the experiment (human vs avatar)

In order to implement the social IOR in our case, we utilized the top-level utterance categories from the speech hierarchy (6 total) so that the Social AI maintained an internal array to keep track of recently spoken categories, with a "counter" that counted down a certain number of seconds during which any further utterances within that same category were suppressed (though the AI could still make utterances from other categories). This counter was set to 3 seconds, based on prior research on social IOR in humans [26]. The study described here involved a control condition, including the social IOR, and an experimental condition where the social IOR was turned off. The total sample size was 32 participants (16 in each condition), with 16 Korean speakers and 16 English speakers split into each condition. Participants played the game with the virtual avatar for 30 minutes total (see Figure 1). To evaluate the effects, we utilized two common standardized scales: Godspeed scale [1] for measuring perceptions of a robot/agent and the Networked Minds instrument [6] for measuring *social presence* [27].

3 RESULTS

Our initial hypothesis was that social IOR would enhance human perception of the interaction with the virtual avatar agent. However, results suggest the effect may be dependent on the language of the speaker. While there was minimal change in the total *social presence* values as per the Networked Minds instrument in Korean speakers without social IOR vs with (0.80 vs 0.85), there was a significant effect in English speakers (0.91 vs. 0.37). This was largely due to increases in both *attentional engagement* (the feeling that when I pay attention to something, the agent does too) and *emotional contagion* (the feeling that when I feel something, the agent feels that way too) in English speakers. Conversely, Korean speakers appeared to have either no change or a slight reduction in those dimensions when social IOR was removed.

Likewise, the Godspeed scores were notably increased on average for English speakers (3.51 vs. 3.06) but not the Korean speakers (3.24 vs. 3.29). In short, whether removing social IOR seems to make the agent more engaging, but less likeable, appears to be dependent on the language of the speaker. We discuss the design implications of this for socially interactive agents further in the next section.

4 DISCUSSION

Social interaction is a notoriously amorphous domain, where it is not always clear how to define the "goals" an agent should pursue [28] and determining the outcome of such goals may be subjective in nature [15]. That leads to challenges with planning for agent behavior, as well as rational decision-making for AI systems in social situations. We contend here that *culturally-appropriate* defiance of social norms can counterintuitively help create autonomous agents and robots that are perceived as more socially intelligent.

This is in line with suggestions from previous research [11, 25], though perhaps more similar to the latter's "intelligent disobedience" approach than the former. In our case, disobedience is not triggered by any specific factors but rather is a design consideration for creating more seemingly intelligent behavior (similar to [16]). **Indeed, a lack of such occasional disobedience may also explain past results in cross-cultural robotics**, which indicate that cultural homophily (e.g. agents adapted to a specific set of cultural attributes) alone does not necessarily correspond to higher ratings of a robot by humans [24]. In certain cultural settings, sporadic purposeful "failures" may actually be a desirable design feature.

Much previous research in HRI has argued for "culturally-robust" or "culturally-aware" systems (including our own), where robots are designed specifically to create adaptable behaviors that match the value system of the local human culture [10, 22, 24, 29]. While that is certainly one approach for *value alignment* in social robots, here we take the position that it may be necessary to create different models (machine learning or otherwise) of robot behavior specifically for different cultures, taking into account differential responses to perceived disobedience in the robot by human interactors. This topic of perceived disobedience towards better social intelligence is an area of rich potential, as there are many ways to violate social norms, which can be dependent on both behavior (speech, facial expressions, gesture) as well as robotic form factor.

ACKNOWLEDGMENTS

This work was supported through funding by a grant from the National Research Foundation of Korea (NRF grant# 2021R1G1A1003801), as well as the research fund of Hanyang University (HY-2020).

REFERENCES

[1] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81.

[2] Casey C. Bennett. 2021. Evoking an intentional stance during human-agent social interaction: Appearances can be deceiving. *IEEE International Symposium on Robot and Human interactive Communication (RO-MAN)* (2021), 362–368.

[3] Casey C. Bennett, Selma Sabanovic, Marlena R. Fraune, and Kate Shaw. 2014. Context congruency and robotic facial expressions: Do effects on human perceptions vary across culture? *IEEE International Symposium on Robot and Human interactive Communication (RO-MAN)* (2014), 465–470.

[4] Casey C. Bennett, Cedomir Stanojevic, Selma Sabanovic, Jennifer A. Piatt, and Seongcheol Kim. 2021. When No One is Watching: Ecological Momentary Assessment to Understand Situated Social Robot Use in Healthcare. *9th International Conference on Human-Agent Interaction (HAI)* (2021), 245–251.

[5] Casey C. Bennett, Benjamin Weiss, Jaeyoung Suh, Eunseo Yoon, Jihong Jeong, and Yejin Chae. 2022. Exploring Data-Driven Components of Socially Intelligent AI through Cooperative Game Paradigms. *Multimodal Technologies and Interaction* 6, 2 (2022), 16.

[6] Frank Biocca, Chad Harms, and Jenn Gregg. 2001. Does language shape thought?: Mandarin and English speakers' conceptions of time. *4th Annual International Workshop on Presence* (2001), 1–9.

[7] Lera Boroditsky. 2001. Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology* 43, 1 (2001), 1–22.

[8] Gordon Briggs, Tom Williams, Ryan Blake Jackson, and Matthias Scheutz. 2021. Why and How Robots Should Say 'No'. *International Journal of Social Robotics* (2021), 1–17.

[9] Thomas J. Bruneau. 1973. Communicative silences: Forms and functions. *Journal of Communication* 23, 1 (1973), 17–46.

[10] Barbara Bruno, Roberto Menicatti, Carmine T. Recchiuto, Edouard Lagrue, Amit K. Pandey, and Antonio Sgorbissa. 2018. Culturally-competent human-robot verbal interaction. *5th International Conference on Ubiquitous Robots (UR)* (2018), 388–395.

[11] Alexandra Coman and David W. Aha. 2018. AI rebel agents. *AI Magazine* 39, 3 (2018), 16–26.

[12] Guy Deutscher. 2010. *Through the language glass: Why the world looks different in other languages.* Metropolitan Books.

[13] Nick J. Enfield. 2017. *How we talk: The inner workings of conversation.* Basic Books.

[14] Orly Fuhrman and Lera Boroditsky. 2010. Cross-cultural differences in mental representations of time: Evidence from an implicit nonlinguistic task. *Cognitive Science* 34, 8 (2010), 1430–1451.

[15] Goren Gordon. 2020. Infant-inspired intrinsically motivated curious robots. *Current Opinion in Behavioral Sciences* 35 (2020), 28–34.

[16] Laura M. Hiatt, Anthony M. Harrison, and J. Gregory Trafton. 2011. Accommodating human variability in human-robot teams through theory of mind. *Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)* (2011).

[17] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology* 9 (2018), 861.

[18] Raymond M. Klein and W. Joseph MacInnes. 1999. Inhibition of return is a foraging facilitator in visual search. *Psychological Science* 10, 4 (1999), 346–352.

[19] Mark L. Knapp, Judith A. Hall, and Terrence G. Horgan. 2013. *Nonverbal communication in human interaction.* Cengage Learning.

[20] Takanori Komatsu, Bertram F. Malle, and Matthias Scheutz. 2021. Blaming the reluctant robot: parallel blame judgments for robots in moral dilemmas across US and Japan. *ACM/IEEE International Conference on Human-Robot Interaction (HRI)* 9 (2021), 63–72.

[21] Stavroula-Thaleia Kousta, David P. Vinson, and Gabriella Vigliocco. 2008. Investigating linguistic relativity through bilingualism: The case of grammatical gender. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34, 4 (2008), 843.

[22] Hee Rin Lee and Selma Sabanovic. 2014. Culturally variable preferences for robot design and use in South Korea, Turkey, and the United States. *ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2014), 17–24.

[23] Velvetina Lim, Maki Rooksby, and Emily S. Cross. 2015. Differences on social acceptance of humanoid robots between Japan and the UK. *Proceedings of the 4th International Symposium on New Frontiers in Human-Robot Interaction, The Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB)* (2015).

[24] Velvetina Lim, Maki Rooksby, and Emily S. Cross. 2020. Social robots on a global stage: establishing a role for culture during human–robot interaction. *International Journal of Social Robotics* (2020), 1–27.

[25] Reuth Mirsky and Peter Stone. 2021. The seeing-eye robot grand challenge: Rethinking automated care. *20th International Conference on Autonomous Agents and MultiAgent Systems* (2021), 28–33.

[26] Orit Nafcha, Simond Shamay-Tsoory, and Shai Gabay. 2020. The sociality of social inhibition of return. *Cognition* 195 (2020), 104108.

[27] Catherine S. Oh, Jeremy N. Bailenson, and Gregory F. Welch. 2018. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI* 5 (2018), 114.

[28] Matthias Rolf and Nigel T. Crook. 2016. What if: Robots create novel goals? Ethics based on social value systems. *EDIA Workshop at the European Conference on Artificial Intelligence (ECAI)* (2016), 20–25.

[29] Selmas Sabanovic, Casey C. Bennett, and Hee-Rin Lee. 2014. Towards culturally robust robots: A critical social perspective on robotics and culture. *Proceedings of the HRI Workshop on Culture-Aware Robotics* (2014).

[30] Benjamin Weiss, Ina Wechsung, Christine Kühnel, and Sebastian Möller. 2015. Evaluating embodied conversational agents in multimodal interfaces. *Computational Cognitive Science* 1, 1 (2015), 1–21.