Toward Understanding the Effects of Socially Aware Robot Behavior

Roesler, Oliver; Bagheri, Elahe; Aly, Amir

Published in:
Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems

DOI:
10.1075/is.22029.roe

Publication date:
2023

Toward understanding the effects of socially aware robot behavior

Oliver Roesler[1], Elahe Bagheri[1], and Amir Aly[2]
[1] IVAI GmbH | [2] University of Plymouth

A key factor for the acceptance of robots as regular partners in human-centered environments is the appropriateness and predictability of their behaviors, which depend partially on the robot behavior's conformity to social norms. Previous experimental studies have shown that robots that follow social norms and the corresponding interactions are perceived more positively by humans than robots or interactions that do not adhere to social norms. However, the conducted studies only focused on the effects of social norm compliance in specific scenarios. Therefore, this paper aims to guide further research studies by compiling how researchers in relevant research fields think the perception of robots and the corresponding interactions are influenced independently of a specific scenario if a robot's behavior conforms to social norms. Additionally, this study investigates what characteristics and metrics constitute a good general benchmark to objectively evaluate the behavior of social robots regarding its conformity to social norms according to researchers in relevant research communities. Finally, the paper summarizes how the obtained results can guide future research toward socially aware robot behavior.

Keywords: social norms, robot behavior adaptation, socially aware robot behavior, metrics, benchmarks

Introduction

The number of robots employed in human-centered environments instead of factories is growing, thereby, increasing the need for robots that can interact with humans in a socially acceptable and appropriate manner. This ranges from socially aware navigation aiming to avoid discomfort (Kruse, Pandey, Alami, & Kirsch, 2013) to socially aware communication, e.g., communicating non-verbally with an adult, if a sleeping child is in the same room (Tomic, Pecora, & Saffiotti, 2014). Due to this vast range of problems that need to be addressed, the work

that aims to integrate social norms, often spans across different disciplines including cognitive robotics, human-robot interaction, and artificial intelligence. For instance, Takayama and Pantofaru (2009) studied the effect of robot gaze on the approach distance humans keep, while Dautenhahn et al. (2006) and Koay, Syrdal, Walters, and Dautenhahn (2007) investigated the appropriate angle by which a robot should approach humans. The findings of these studies show that it is possible to improve humans' impression of robots by enabling them to follow human social norms. Ciou, Hsiao, Wu, Tseng, and Fu (2018) showed that the movements of robots that follow social norms are perceived as smoother, more predictable, and rarely disturbing. Important to note is that some social norms differ across cultures, such as greeting and speaking language, and that people prefer robots that follow the social norms of the culture they are belonging to (Trovato et al., 2013; Wang, Rau, Evers, Robinson, & Hinds, 2010). Further, Wang et al. (2010) showed that when a robot shows the same culture, the participants are more likely to change their decisions to align with the robot.

Overall, several previous studies showed that following social norms improves the perception of – and interaction with – robots resulting in improved user experience and acceptance of social robots (Joosse, Lohse, & Evers, 2014). In addition, robots that are not able to follow social norms are likely to violate user expectations, which cannot only lead to impoverished interactions but in some cases even to emotional or physical harm (Sarathy, Wilson, Arnold, & Scheutz, 2016). Although endowing robots with the ability to follow social norms is beneficial, it is a challenging task due to the large number of existing social norms, their dynamicity (some norms change over time), and strong variance across different social and cultural groups. Thus, the question arises whether social norms can be manually defined and hard-coded into the robot controllers or whether they need to be learned automatically through interaction?

Hard-coding social norms in robots is commonly used and has been useful to prove the benefits of norm conformity by comparing the behavior generated by a model that takes social norms into account with a baseline model without any information about social norms, e.g., (Dautenhahn et al., 2006; Koay et al., 2007; Takayama & Pantofaru, 2009). However, these models usually work only for the specific scenarios encountered during the conducted experiment and would require a lot of time and effort to be tuned for other scenarios, thus, hard-coding socially normative behavior is not scalable. In contrast, employing machine learning to learn socially normative behavior can reduce the time and effort to learn the appropriate behavior for different situations; however, to achieve good results, a decent amount of data is necessary that requires conducting many interactions, which is not necessarily a trivial target. Nevertheless, different learning approaches, including reinforcement learning and deep learning-based models,

have recently been used to enable robots to adapt their behaviors to social norms. For instance, Ciou et al. (2018) and Gao et al. (2019) used deep reinforcement learning to teach robots social navigation and appropriate approaching behavior based on sensor input and human feedback.

Finally, to be able to objectively compare different approaches and models (independent of whether they have been hard-coded or learned), it is important to define general benchmarks and evaluation metrics. However, this is non-trivial and has; therefore, despite its importance, not received much attention. To define a general benchmark, different parameters and factors should be considered, e.g., whether robots should follow the same social norms as humans or develop their own. The latter might be necessary because people might not like robots to be treated the same as humans, especially when considering non-humanoid robots, but this might strongly depend on culture. For example, Li, Rau, and Li (2010) discovered that participants from low-context cultures, like German, rated robot behavior differently in respect to its conformity to social norms than those from high-context cultures, like Chinese or Korean.

In this paper, we present a subjective analysis to foster a better understanding of what researchers in relevant research communities think is the effect of norm conformity on the perception of a robot's behavior and the interaction. Furthermore, the analysis aims to determine what characteristics and metrics they think constitute a good general benchmark to support the objective evaluation of socially normative robot behaviors. The former is important to identify views that have not yet been verified through experiments to ensure that future research is not based on wrong assumptions, while the latter provides a good starting point to create a general benchmark. The analysis utilizes three separate parts: (1) an online survey, (2) an interactive session, and (3) a panel discussion.[1] The remainder of the paper is structured as follows: First, we present a detailed analysis of the collected online survey data. Afterwards, we summarize the main points made during the interactive session and panel discussion. Finally, we conclude the paper by highlighting the key results and discussing their implications for future research in the area of socially aware robot behavior adaptation.

Survey

We created an online survey to compile an overview of how socially aware robot behavior, which conforms to social norms, will influence the perception of robots

1. The interactive session and panel discussion were held at the TSAR 2021 workshop in August 2021 (https://tsar2021.ai.vub.ac.be/).

by humans who are interacting with them and to determine which characteristics constitute a good benchmark for the evaluation of social robot behavior regarding its compliance to social norms. All statements that were part of the survey are listed in Tables (4 and 5). The survey was initially distributed among the participants of the virtual workshop "Robot Behavior Adaptation to Human Social Norms" on August 12, 2021, however, due to the virtual format of the survey, we sent its URL to other colleagues via relevant mailing lists to ensure a large number of responses. Overall, a total of 109 researchers (52 female, 56 male, and 1 other) completed the survey.[2] We do not claim that the collected responses are representative of the relevant research communities as a whole, however, they provide a good indication of the important topics, trends, and open research questions when considering the deployment of social robots in human-centered environments in general and more specifically, the development of mechanisms for the dynamic adaptation of robot behaviors to social norms.

Demographics

Both large robotics conferences, like IEEE RO-MAN, and English-speaking online mailing lists are usually very diverse with respect to the countries of origin of the participants, i.e., where the participants grew up, and the countries of residence, i.e., where the participants are currently living. Nevertheless, specific research topics sometimes receive larger attention in specific geographic locations, which is why we asked the participants of the survey for their countries of origin and residence to ensure awareness of potential geographical or cultural biases. Figure (1a) shows that people from all ten different cultural clusters[3] participated in the study with the largest group being people who grew up in countries belonging to Latin-European cultures and the smallest group people who grew up in African, Latin-American, and Nordic cultures. When looking at the countries of residence, we can see a shift from South-East Asian, Middle Eastern, and Eastern European cultures toward Anglo-Saxon and Nordic cultures (Figure 1b). The gender of the participants was well distributed as 51.4% for males and 47.7% for females (Figure 1c). Figure (1d) shows that about 66.1% of the participants came from academia. Additionally, 73.4% reported their primary field of research as Human-Robot Interaction (HRI), Artificial Intelligence (AI), or robotics. Fig-

2. All participants provided their consent before starting the survey.

3. The countries provided by the participants have been grouped into the ten different cultural clusters defined by Mensah and Chen (2013).

ure (1f) shows that most of the participants, i.e., 69.4%, were under 41 years old, and 25% were between 41 to 61 years old.

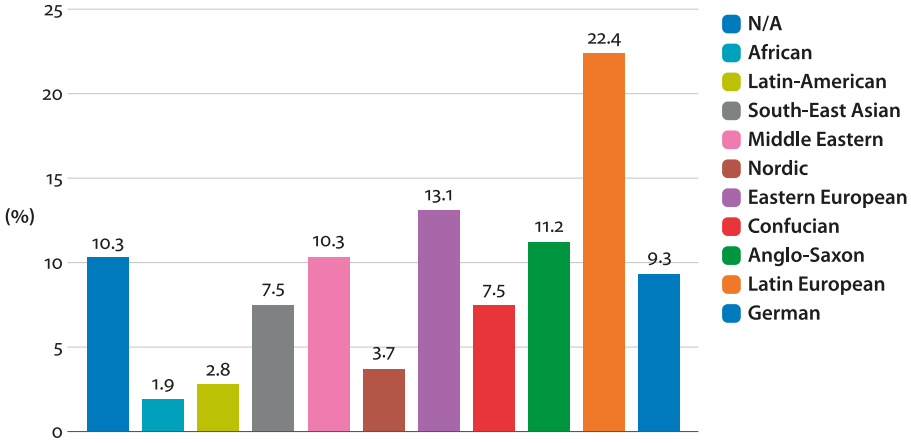
Statistical analysis

To verify whether any significant differences exist based on the participants' gender, age, education, occupation, primary field of research, country of origin (i.e., where a person grew up), or country of residence (i.e., where a person is currently living), several statistical tests were applied. Since multiple similar questions were shown together in the survey (G01-G10), we first applied a Cronbach's alpha test to verify the reliability of the statements in each group (Table 1). The obtained results show that the results are reliable for seven of the ten groups, i.e., G01-G04, G06, G09, and G10, when considering α as 0.60.

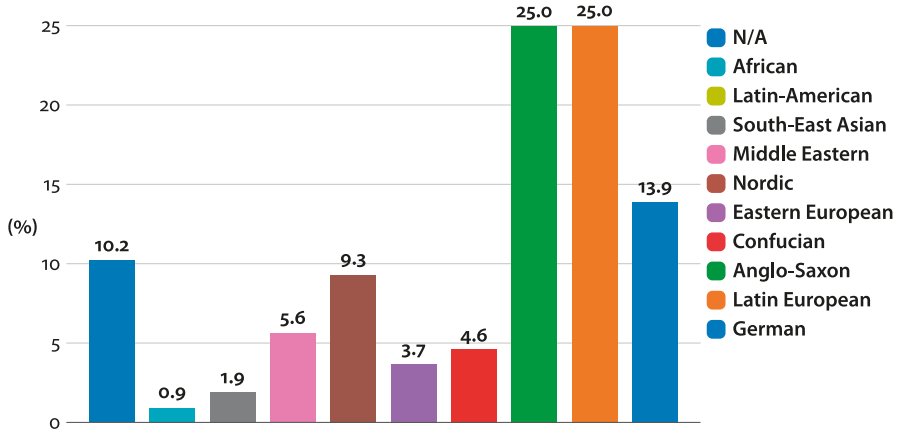
Table 1. Results of the Cronbach's test for the 10 different statement groups. Since α is set to 0.60, the obtained results are reliable for seven groups (shown in bold). The statements belonging to each group are listed in Tables (4 and 5)

Groups	α	N
G01	**0.85**	6
G02	**0.77**	5
G03	**0.82**	3
G04	**0.78**	9
G05	0.48	4
G06	**0.74**	5
G07	0.58	4
G08	0.41	4
G09	**0.76**	8
G10	**0.61**	4

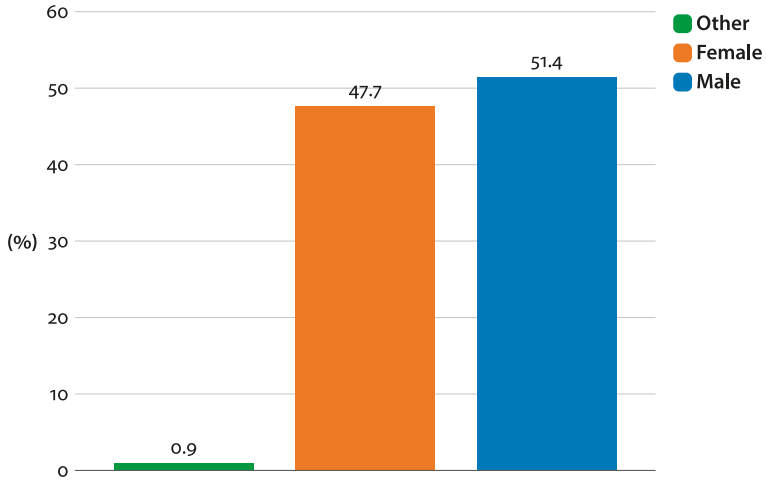
Since the data is not normal, which was determined by applying the Shapiro-Francia test, we used an independent T-test for gender, since we excluded the "other" category from the analysis because only one participant (<1%) indicated its gender as "other", and Kruskal-Wallis tests for non-binary classes, i.e., age, primary field of research, occupation, country of origin, and country of residence. Groups with fewer than 5 samples were discarded. For the age analysis, only the participants under 60 years were considered because the number of participants above 60 years was less than 5 for each group (the participants were grouped in 10



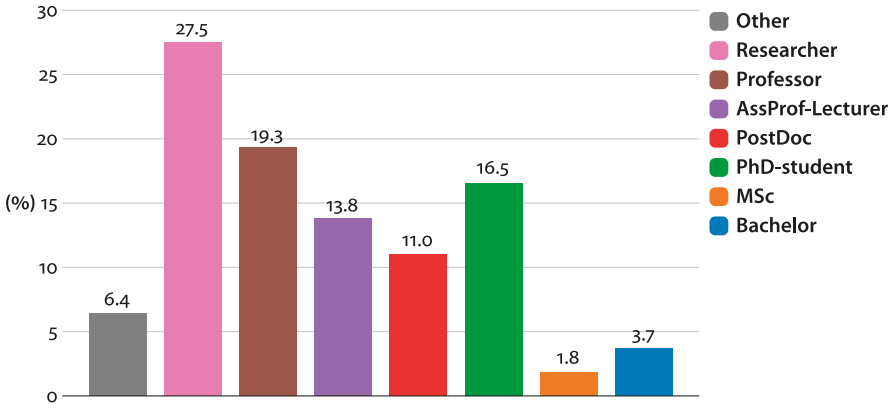
a. The country of origin of the participants according to cultural clusters (Mensah & Chen, 2013)



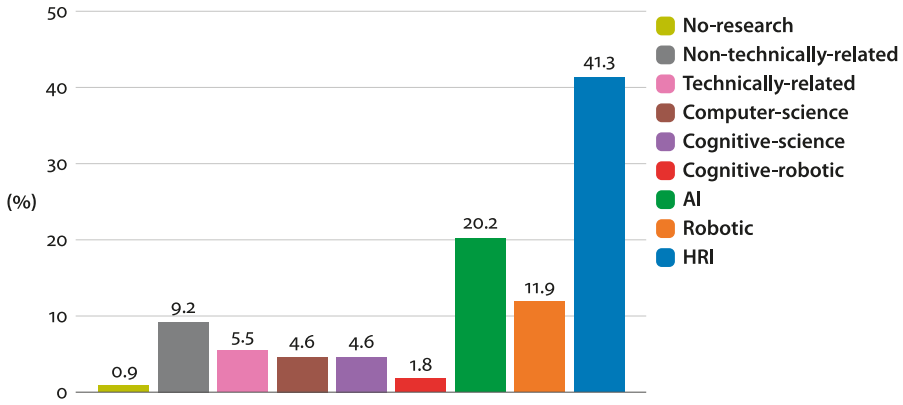
b. The country of residence of the participants according to cultural clusters (Mensah & Chen, 2013)



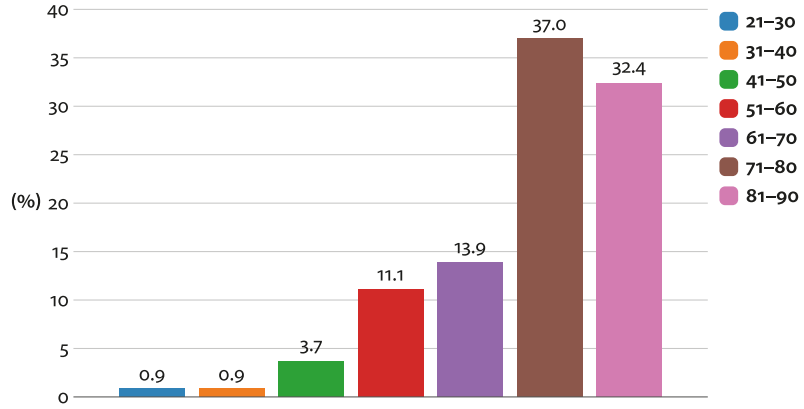
c. The gender of the participants



d. The occupation of the participants



e. The research fields that the participants reported to belong to



f. The age ranges (in years) of the participants

Figure 1. Overview of the demographics of the participants in the survey. For 11 participants no information about their country of origin and country of residence was provided (shown as N/A)

years intervals). Since the Kruskal-Wallis tests only show that there are statistically significant differences between groups without specifying the specific groups, the Mann-Whitney U test was applied for all statements that had p-values below 0.05 to identify the groups with statistically significant difference. We report results both without and with Bonferroni correction because the former can lead to false positives (Type I error) and the latter to false negatives (Type II error) as dis-

cussed by Cabin and Mitchell (2000); Nakagawa (2004); Perneger (1998) among others.

Gender

The results of the independent T-test (Table 2) show that there is a significant difference based on gender for the following five statements:

- S27 ($P = 0.03$, $M_f = 5.11$, $SD_f = 1.4$, $M_m = 4.5$, $SD_m = 1.5$)
- S33 ($P = 0.01$, $M_f = 5.63$, $SD_f = 1.33$, $M_m = 4.98$, $SD_m = 1.43$)
- S35 ($P = 0.01$, $M_f = 4.92$, $SD_f = 1.06$, $M_m = 5.46$, $SD_m = 1.14$)
- G06 ($P = 0.02$, $M_f = 6.18$, $SD_f = 0.52$, $M_m = 5.9$, $SD_m = 0.72$)
- G09 ($P = 0.02$, $M_f = 5.68$, $SD_f = 0.68$, $M_m = 5.39$, $SD_m = 0.73$)

Age

The results of the statistical analysis (Table 2) show that there is a significant difference based on age for S43 ($P = 0.04$) and S45 ($P = 0.04$). More specifically, while the participants who are 21–40 years old and those who are 41–50 years old both overall agree that "S43: Scenarios should allow active/online learning of normative behavior", the latter (i.e., 41–50 years old) only agree slightly with it. Similarly, the older participants who are 41–60 years old agree that "S45: Robots should be evaluated regarding their adaptability to different users"; however, their agreement is significantly less than that of the younger participants who are 21–30 years old. However, after applying a Bonferroni correction to counteract the multiple comparisons problem, there are no more significant differences between groups (Table 3).

Primary field of research

Table (2) shows that there are no significant differences based on the primary field of research reported by the participants.

Occupation

The results in Table (2) show that there is a significant difference based on the participant occupation for G02 ($P = 0.03$) and S43 ($P = 0.04$). More specifically, professors, researchers, and "others" agreed that "Interaction with robots that follow social norms will be more predictable, enjoyful, efficient, natural, and comfortable" (G02), while assistant professors and post-doctoral associates only slightly agreed with these statements. For S43, i.e., "Scenarios should allow active/online learning of normative behavior", professors, researchers, and post-doctoral associates agreed with it, while assistant professors agreed only slightly with it. Possible reasons for this difference might be due to the participant's age since most post-

doctoral associates, assistant professors, and lecturers who participated in the survey were 41–50 years old, while most professors were more than 60 years old or a difference between academia (post-doctoral associates, assistant professors, lecturers, and professors) and non-academic groups (researchers and "others"). However, after performing a Bonferroni correction to control for the family-wise error rate, the differences are no longer statistically significant (Table 3).

Country of origin

For the country of origin there are significant differences in the participants' rating for five statements: S25 ($P = 0.01$), S29 ($P = 0.00$), S34 ($P = 0.00$), S46 ($P = 0.02$), and S51 ($P = 0.00$) (Table 2). For S25, i.e., "The behavior of a robot is more important than its appearance for entertainment", and S34, i.e., "Sharing experiences through cloud-based robot systems will lead to less personalized social behaviors", there are statistically significant differences between cultural cluster 8 (CC8) and CC1-CC5, and between CC8 and CC7 (Table 3). For S25, the participants belonging to CC8 (South-East Asian) strongly agree with the statement, while the participants belonging to CC2 (Latin European), CC3 (Anglo-Saxon), CC4 (Confucian), and CC7 (Middle Eastern) agree, or at least slightly agree, with it. The participants belonging to CC1 (German) and CC5 (Eastern European) are mostly neutral about it, and the participants belonging to CC1 are even inclined to slightly disagree with it. Additionally, there is a statistically significant difference between CC1 and CC2, and between CC1 and CC4 for S25. For S34, the participants belonging to CC8 agree with the statement, while the participants belonging to CC1, CC3, CC5, and CC7 are slightly disagreeing with it, and the participants belonging to CC2 and CC4 are neither agreeing nor disagreeing with it. For S29, i.e., "Socially normative robot behavior cannot be hard-coded or programmed but must be learned through interaction with humans", there are statistically significant differences between CC1, CC2, CC3, and CC5 and CC8 as well as between CC3, and CC5 and CC4 with the participants belonging to CC4 and CC8 agreeing with the statement, while the participants belonging to CC1 and CC3 are neutral about it, and the participants belonging to CC2 and CC5 are slightly agreeing or disagreeing with it, respectively. For S46, i.e., "Robots should be evaluated regarding their adaptability to different environments.", the participants belonging to CC2 and CC8 rated the statement significantly different from the participants belonging to CC5 with the former agreeing or even strongly agreeing with it and the latter only slightly agreeing with it. Additionally, there is a statistically significant difference between CC3 and CC8 with the former agreeing less strongly than the latter with S46. Finally, for S51, i.e., "The interaction should be evaluated regarding its success (does the robot achieve its goal?)", there are statistically significant differences between CC1 and CC2, CC4, CC7, as well

as CC8 with the first being neutral and the rest agreeing with it. Similarly, the participants belonging to CC3 and CC5 are between neutral and slight agreement in contrast to CC4 and CC8. Overall, the most differences are for CC8 (South-East Asian), and in most cases, the participants belonging to CC8 were more strongly agreeing with the statements. In contrast, CC1 (German), which was the cultural cluster with the second-highest number of statistically significant differences, mostly agreed less or disagreed more with the five statements than the participants belonging to the other cultural clusters. However, after applying a Bonferroni correction there is only a significant difference for S34 between CC5 (Eastern European) and CC8 (South-East Asian) ($U = 5.5$, $P = 0.0014$). Where the participants belonging to CC8 agree with the statement ($M = 6$, $SD = 0.75$), while the participants belonging to CC5 are slightly disagreeing with it ($M = 3.84$, $SD = 1.02$) (Table 3).

Table 2. Results of an independent T-test for the obtained data based on the participants' gender and Kruskal-Wallis tests for the obtained data based on the participants' age, primary field of research, occupation, country of origin, and country of residence. The statements are explained in Tables (4 and 5)

	Gender	Age	Field	Occupation	Origin	Residence
G01	0.94	0.08	0.29	0.37	0.06	0.92
G02	0.73	0.72	0.36	**0.03**	0.74	0.55
G03	0.83	0.40	0.49	0.29	0.64	0.54
G04	0.99	0.70	0.69	0.49	0.46	0.32
S24	0.09	0.13	0.48	0.11	0.57	0.60
S25	0.06	0.97	0.16	0.17	**0.01**	0.30
S26	0.37	0.28	0.28	0.65	0.78	0.33
S27	**0.03**	0.91	0.52	0.49	0.46	0.61
S28	0.69	0.38	0.55	0.70	0.22	0.50
S29	0.44	0.41	0.45	0.31	**0.00**	**0.00**
S30	0.88	0.53	0.44	0.67	0.18	0.19
S31	0.31	0.89	0.72	0.76	0.17	0.44
S32	0.19	0.88	0.39	0.98	0.78	0.76
S33	**0.01**	0.08	0.64	0.09	0.17	0.48
S34	0.74	0.99	0.77	0.85	**0.00**	0.50
S35	**0.01**	0.46	0.21	0.80	0.19	0.22
S36	0.71	0.59	0.87	0.71	0.39	0.49

Table 2. *(continued)*

	Gender	Age	Field	Occupation	Origin	Residence
G06	**0.02**	0.32	0.18	0.24	0.11	0.12
S42	0.38	0.09	0.55	0.08	0.06	0.53
S43	0.74	**0.04**	0.95	**0.04**	0.49	0.32
S44	0.81	0.29	0.35	0.14	0.05	0.05
S45	0.63	**0.04**	0.50	0.09	0.24	0.24
S46	0.84	0.17	0.18	0.51	**0.02**	0.39
S47	0.78	0.96	0.53	0.56	0.17	0.13
S48	0.51	0.24	0.13	0.94	0.09	0.05
S49	0.14	0.46	0.94	0.89	0.10	0.85
S50	0.13	0.13	0.88	0.67	0.50	0.98
S51	0.66	0.28	0.91	0.93	**0.00**	**0.01**
G09	**0.04**	0.76	0.53	0.26	0.30	0.41
G10	0.25	0.29	0.74	0.06	0.23	0.47

Table 3. Results of a Mann-Whitney U test for all the statements that showed a statistically significant difference ($P < 0.05$) based on their Kruskal-Wallis tests (Table 2). The statements are explained in Tables (4 and 5). AC, OC, and CC refer to age cluster, occupation cluster, and cultural cluster, respectively. After applying Bonferroni corrections a statistically significant difference was only shown for CC5 and CC8 for statement S34 (shown in bold)

			Mann-Whitney U		Bonferroni
	Statement	**Groups**	**U**	**P**	***a***
Age	S43	AC3 ($M = 5.33$, $SD = 0.62$) and AC1 ($M = 5.82$, $SD = 1.17$)	283.5	0.0366	0.0083
		AC3 ($M = 5.33$, $SD = 0.62$) and AC2 ($M = 5.97$, $SD = 0.99$)	344.0	0.0089	
	S45	AC3 ($M = 5.5$, $SD = 0.5$) and AC1 ($M = 6.0$, $SD = 0.80$)	285.0	0.0277	
		AC4 ($M = 5.23$, $SD = 1.18$) and AC1 ($M = 6.0$, $SD = 0.80$)	302.0	0.0381	
Occupation	G02	OC5 ($M = 5.27$, $SD = 0.7$) and OC9 ($M = 6$, $SD = 0.79$)	12.5	0.0427	0.0033

Table 3. *(continued)*

			Mann-Whitney U		Bonferroni
	Statement	**Groups**	**U**	**P**	***a***
		OC6 (*M = 4.75, SD = 1.19*) and OC7 (*M = 5.68 SD = 0.57*)	73.5	0.0198	
		OC6 (*M = 4.75, SD = 1.19*) and OC8 (*M = 5.76, SD = 0.71*)	99.0	0.0070	
		OC6 (*M = 4.75, SD = 1.19*) and OC9 (*M = 6, SD = 0.79*)	13.5	0.0203	
	S43	OC6 (*M = 4.85, SD = 1.3*) and OC5 (*M = 6.27, SD = 0.61*)	123.0	0.0098	
		OC6 (*M = 4.85, SD = 1.3*) and OC7 (*M = 5.75, SD = 0.82*)	84.5	0.0408	
		OC6 (*M = 4.85, SD = 1.3*) and OC8 (*M = 6, SD = 0.74*)	102.5	0.0062	
Origin	S25	CC1 (*M = 3.88, SD = 1.44*) and CC2 (*M = 5.17, SD = 1.73*)	56.0	0.0445	0.0018
		CC1 (*M = 3.88, SD = 1.44*) and CC4 (*M = 5.85, SD = 0.83*)	9.0	0.0145	
		CC1 (*M = 3.88, SD = 1.44*) and CC8 (*M = 6.57, SD = 1.04*)	4.0	0.0031	
		CC2 (*M = 5.17, SD = 1.73*) and CC8 (*M = 6.57, SD = 1.04*)	34.0	0.0193	
		CC3 (*M = 5.36, SD = 1.61*) and CC8 (*M = 6.57, SD = 1.04*)	17.5	0.0458	
		CC4 (*M = 5.85, SD = 0.83*) and CC8 (*M = 6.57, SD = 1.04*)	9.5	0.0420	
		CC5 (*M = 4.3, SD = 1.85*) and CC8 (*M = 6.57, SD = 1.04*)	13.5	0.0088	
		CC7 (*M = 5.3, SD = 1.48*) and CC8 (*M = 6.57, SD = 1.04*)	14.0	0.0329	
	S29	CC3 (*M = 3.9, SD = 1.62*) and CC4 (*M = 5.71, SD = 0.45*)	9.5	0.0066	
		CC5 (*M = 3.76, SD = 1.62*) and CC4 (*M = 5.71, SD = 0.45*)	77.5	0.0108	
		CC1 (*M = 4.11, SD = 1.79*) and CC8 (*M = 6.28, SD = 0.88*)	9.5	0.0198	

Table 3. *(continued)*

			Mann-Whitney U		Bonferroni
	Statement	**Groups**	**U**	**P**	***a***
		CC2 (*M = 4.73*, *SD = 1.48*) and CC8 (*M = 6.28*, *SD = 0.88*)	32.5	0.0173	
		CC3 (*M = 3.9*, *SD = 1.62*) and CC8 (*M = 6.28*, *SD = 0.88*)	7.5	0.0041	
		CC5 (*M = 3.76*, *SD = 1.62*) and CC8 (*M = 6.28*, *SD = 0.88*)	9.5	0.0042	
	S34	CC1 (*M = 3.11*, *SD = 1.44*) and CC8 (*M = 6*, *SD = 0.75*)	3.5	0.0031	
		CC2 (*M = 3.95*, *SD = 1.51*) and CC8 (*M = 6*, *SD = 0.75*)	22.0	0.0038	
		CC3 (*M = 3.09*, *SD = 1.67*) and CC8 (*M = 6*, *SD = 0.75*)	4.5	0.0020	
		CC4 (*M = 4.28*, *SD = 1.27*) and CC8 (*M = 6*, *SD = 0.75*)	7.0	0.0268	
		CC5 (*M = 3.48*, *SD = 1.02*) and CC8 (*M = 6*, *SD = 0.75*)	**5.5**	**0.0014**	
		CC7 (*M = 3.1*, *SD = 1.51*) and CC8 (*M = 6*, *SD = 0.75*)	4.5	0.0290	
	S46	CC5 (*M = 4.92*, *SD = 1.54*) and CC2 (*M = 6.04*, *SD = 0.95*)	212.0	0.0341	
		CC3 (*M = 5.81*, *SD = 0.57*) and CC8 (*M = 6.57*, *SD = 0.72*)	16.5	0.0364	
		CC5 (*M = 4.92*, *SD = 1.54*) and CC8 (*M = 6.57*, *SD = 0.72*)	16.0	0.0176	
	S51	CC1 (*M = 4.11*, *SD = 0.87*) and CC2 (*M = 5.08*, *SD = 1.05*)	46.0	0.0126	
		CC1 (*M = 4.11*, *SD = 0.87*) and CC4 (*M = 5.85*, *SD = 0.63*)	5.0	0.0043	
		CC1 (*M = 4.11*, *SD = 0.87*) and CC7 (*M = 5.7*, *SD = 1.1*)	13.5	0.0088	
		CC1 (*M = 4.11*, *SD = 0.87*) and CC8 (*M = 5.85*, *SD = 1.35*)	10.5	0.0261	
		CC3 (*M = 4.63*, *SD = 1.14*) and CC4 (*M = 5.85*, *SD = 0.63*)	14.5	0.0279	

Table 3. *(continued)*

			Mann-Whitney U		Bonferroni
	Statement	**Groups**	**U**	**P**	***a***
		CC5 ($M = 4.76$, $SD = 0.79$) and CC4 ($M = 5.85$, $SD = 0.63$)	75.0	0.0157	
		CC5 ($M = 4.76$, $SD = 0.79$) and CC8 ($M = 5.85$, $SD = 1.35$)	21.0	0.0498	
Residence	S29	CC1 ($M = 4.14$, $SD = 1.64$) and CC4 ($M = 6.25$, $SD = 0.82$)	8.0	0.0346	0.0033
		CC1 ($M = 4.14$, $SD = 1.64$) and CC7 ($M = 6$, $SD = 0.63$)	11.5	0.0293	
		CC3 ($M = 3.96$, $SD = 1.5$) and CC2 ($M = 5.07$, $SD = 1.43$)	474.0	0.0111	
		CC3 ($M = 3.96$, $SD = 1.5$) and CC4 ($M = 6.25$, $SD = 0.82$)	11.0	0.0119	
		CC3 ($M = 3.96$, $SD = 1.5$) and CC7 ($M = 6$, $SD = 0.63$)	16.0	0.0078	
	S51	CC2 ($M = 5.19$, $SD = 0.78$) and CC5 ($M = 6.33$, $SD = 0.47$)	10.5	0.0329	
		CC3 ($M = 4.84$, $SD = 1.19$) and CC5 ($M = 6.33$, $SD = 0.47$)	10.5	0.0375	

Table 4. List of the used statements regarding the influence of socially normative behavior on the perception of the robot and interaction. The statements S28-S35 are separate statements that do not belong to any group

Group	**Code**	**Statement**
G01	S01	Robots that follow social norms will be perceived as more friendly.
	S02	Robots that follow social norms will be perceived as more human-like.
	S03	Robots that follow social norms will be perceived as more empathic.
	S04	Robots that follow social norms will be perceived as more understanding.
	S05	Robots that follow social norms will be perceived as more predictable.
	S06	Robots that follow social norms will be perceived as more trustworthy.
G02	S07	Interactions with robots that follow social norms will be more predictable.
	S08	Interactions with robots that follow social norms will be more enjoyful.
	S09	Interactions with robots that follow social norms will be more efficient.

Table 4. *(continued)*

Group	Code	Statement
	S10	Interactions with robots that follow social norms will be more natural.
	S11	Interactions with robots that follow social norms will be more comfortable.
G03	S12	Compliance with social norms will lead to greater user satisfaction.
	S13	Compliance with social norms will lead to longer interactions.
	S14	Compliance with social norms will lead to stronger human-robots relationships.
G04	S15	Appropriate robot behavior depends on the robot's perceived gender including genderless.
	S16	Appropriate robot behavior depends on the relationship between the human and robot, e.g., due to previous interactions.
	S17	Appropriate robot behavior depends on the human's prior experience with robots.
	S18	Appropriate robot behavior depends on the human's personality.
	S19	Appropriate robot behavior depends on the human's gender.
	S20	Appropriate robot behavior depends on the human's age.
	S21	Appropriate robot behavior depends on the human's culture.
	S22	Appropriate robot behavior depends on the robot's type and shape, e.g., humanoids or drones.
	S23	Appropriate robot behavior depends on the area of application, e.g., healthcare entertainment etc.
G05	S24	The behavior of a robot is more important than its appearance for education.
	S25	The behavior of a robot is more important than its appearance for entertainment.
	S26	The behavior of a robot is more important than its appearance for drone-based applications.
	S27	The behavior of a robot is more important than its appearance for health/ elderly care.
	S28	The social norms that robots should follow are different from human social norms.
	S29	Socially normative robot behavior cannot be hard-coded or programmed but must be learned through interaction with humans.
	S30	Appropriate robot behavior cannot solely be learned from watching human-human interaction but must be learned through interaction with humans.
	S31	Following social norms, i.e., what the group believes is appropriate, is more important than taking care of an individual's personal preferences.

Table 4. *(continued)*

Group	Code	Statement
	S32	Evaluation of robot behavior regarding its conformity to social norms must be done in natural human environments and cannot be done in artificial laboratory settings.
	S33	Allowing robots to take into account the descriptive characteristics of a user, e.g., appearance, to determine appropriate socially normative behaviors can lead to discrimination and racism.
	S34	Sharing experiences through cloud-based robot systems will lead to less personalized social behaviors.
	S35	Social norms can be learned under conditions of uncertainty.

Table 5. List of the used statements regarding benchmarks and metrics to evaluate socially normative robot behavior. The statements S36 and S42-S43 are separate statements that do not belong to any group

Group	Code	Statement
	S36	The benchmark should consist of multiple scenarios that can be independently used.
G06	S37	The following scenarios/problems should be included: interaction in different social settings, e.g., teacher-student, guide, social companion, etc.
	S38	The following scenarios/problems should be included: interaction with different cultures and age groups.
	S39	The following scenarios/problems should be included: multi-party human-robot interaction.
	S40	The following scenarios/problems should be included: physical human-robot collaboration.
	S41	The following scenarios/problems should be included: socially aware navigation (proxemics).
	S42	The scenarios should be independent of a specific robot, like the RoboCup Open Platform League, to allow investigation of the influence of shape and design.
	S43	Scenarios should allow active/online learning of normative behavior.
G07	S44	Robots should be evaluated regarding their safety.
	S45	Robots should be evaluated regarding their adaptability to different users.
	S46	Robots should be evaluated regarding their adaptability to different environments.

Table 5. *(continued)*

Group	Code	Statement
	S47	Robots should be evaluated regarding their ability to handle multiple interactions simultaneously.
G08	S48	The interaction should be evaluated regarding its similarity to human-human interaction.
	S49	The interaction should be evaluated regarding its amount of engagement of the human(s).
	S50	The interaction should be evaluated regarding its efficiency.
	S51	The interaction should be evaluated regarding its success (does the robot achieve its goal?).
G09	S52	Robots should be evaluated regarding their perceived self-awareness.
	S53	Robots should be evaluated regarding their perceived human-awareness.
	S54	Robots should be evaluated regarding their perceived predictability.
	S55	Robots should be evaluated regarding their perceived human-likeness.
	S56	Robots should be evaluated regarding their perceived friendliness.
	S57	Robots should be evaluated regarding their perceived trustworthiness.
	S58	Robots should be evaluated regarding their perceived safety.
	S59	Robots should be evaluated regarding their perceived flexibility.
G10	S60	The social interaction should be evaluated regarding its perceived comfortability.
	S61	The social interaction should be evaluated regarding its perceived naturalness.
	S62	The social interaction should be evaluated regarding its perceived efficiency.
	S63	The social interaction should be evaluated regarding its perceived predictability.

Country of residence

There are less statistically significant differences between the participants' ratings based on the country of residence with respect to the country of origin. One reason might be that CC8 (South-East Asian), which had the largest number of differences when considering the country of origin, is not considered for the statistical analysis for the country of residence because it has only 2 participants. Nevertheless, Table (2) shows that there are significant differences for S29 ($P=0.00$) and S51 ($P=0.01$), which also showed significant differences for the country of origin. For S29, i.e., "Socially normative robot behavior cannot be hard-coded or programmed but must be learned through interaction with humans", the participants belonging to CC1 (German) and CC3 (Anglo-Saxon)

were mostly neutral about it, while the participants belonging to CC2 (Latin European) slightly agreed, and the participants belonging to CC4 (Confucian) and CC5 (Eastern European) agreed with the statement. For S51, i.e., "The interaction should be evaluated regarding its success (does the robot achieve its goal?)", the participants belonging to CC5 agreed with it, while the participants belonging to CC2 and CC3 only slightly agreed with it. Interesting is that the participants who grew up in Easter European countries (CC5) only slightly agreed with S51, while the participants who currently live in countries belonging to CC5 are agreeing with it. However, after performing a Bonferroni correction, none of the differences between groups are statistically significant (Table 3).

Influence of socially normative behavior on the perception of the robot and interaction

We were interested to see how the participants think the perception of a robot and the interaction will change, when the robot's behavior follows social norms, and how much importance participants contribute to following social norms when it comes to the acceptance of social robots as human partners. Therefore, we asked the participants to rate 35 statements (Table 4) covering the described topics on a 7-point Likert scale ranging from "strongly disagree" to "strongly agree".

First, we asked the participants how they think the perception of robot behaviors would change when they follow social norms. Figure 2 shows that the participants think that robots – following social norms – are perceived as more trustworthy ($M=5.69$, $SD=1.24$), human-like ($M=5.55$, $SD=1.14$), empathic ($M=5.52$, $SD=1.28$), friendly ($M=5.36$, $SD=1.25$), predictable ($M=5.22$, $SD=1.34$), and understanding ($M=4.93$, $SD=1.32$). Afterwards, we asked the participants how they think interactions would change in case robots would behave according to social norms. Figure 3 shows that the participants mostly think that interactions would be more predictable ($M=5.74$, $SD=1.2$), enjoyful ($M=5.71$, $SD=1.26$), comfortable ($M=5.47$, $SD=1.33$), natural ($M=5.4$, $SD=1.21$), and efficient ($M=5.32$, $SD=1.31$). Finally, we asked the participants regarding the effect of robot behavior conformity to social norms on user satisfaction, interaction length, and human-robot relationships. Figure 4 shows that the participants believe that compliance with social norms will lead to greater user satisfaction (S12: $M=5.28$, $SD=1.2$), longer interactions (S13: $M=5.43$, $SD=1.23$), and stronger human-robot relationships (S14: $M=5.66$, $SD=1.17$). Overall, the participants think that humans will have better interaction with robots that follow social norms both from an emotional perspective as well as in terms of efficiency.

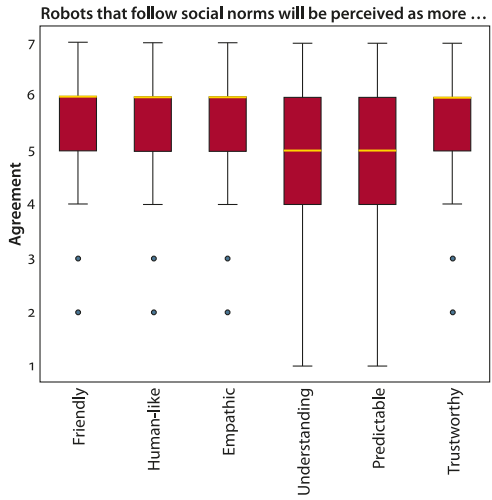


Figure 2. Medians and standard deviations for statements aiming to determine how the compliance to social norms affects how robots are perceived

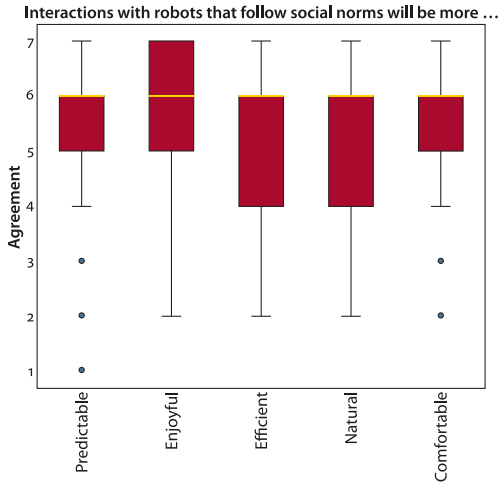


Figure 3. Medians and standard deviations for statements aiming to determine how the compliance to social norms affects human-robot interactions

Figure 5 shows the parameters that can have an effect on whether a specific robot behavior is seen as being appropriate or not. Based on the participants' ratings, there was no agreement about whether human's age (S20: $M = 4.16$, $SD = 1.72$) and prior experience with robots (S17: $M = 4.18$, $SD = 1.64$) have an influence on seeing the robot's behavior as being appropriate. In contrast, the participants

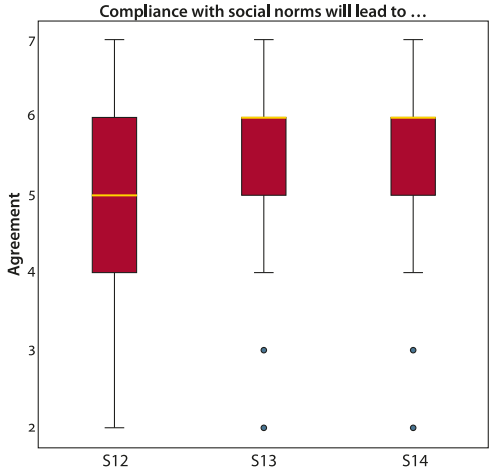


Figure 4. Medians and standard deviations for statements aiming to determine effects of social norm compliance on the users and interaction. The complete statements are provided in Table (4)

agreed that the relationship between a human and a robot (S16: $M = 5.86$, $SD = 1.45$), human's personality (S18: $M = 5.82$, $SD = 1.2$), area of application (S23: $M = 5.65$, $SD = 1.21$), human's culture (S21: $M = 5.46$, $SD = 1.17$), robot's type and shape (S22: $M = 5.45$, $SD = 1.2$), robot's perceived gender (S15: $M = 5.44$, $SD = 1.61$), and human's gender (S19: $M = 5.35$, $SD = 1.25$) have an influence on whether a specific robot behavior is seen as being appropriate.

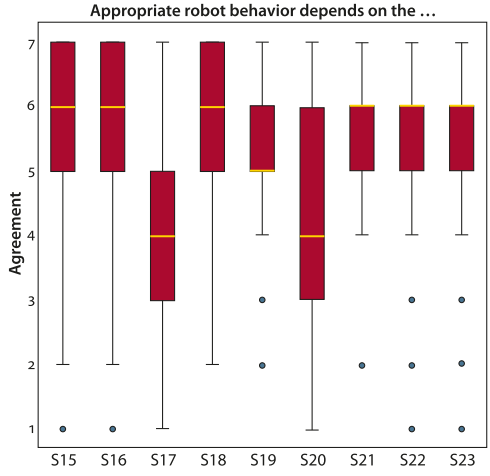


Figure 5. Medians and standard deviations for statements aiming to determine which factors could affect a specific robot behavior to be seen appropriate. The complete statements are provided in Table (4)

Moreover, we asked the participants to determine whether the robot behavior is more important than its appearance for different contexts. Figure 6 shows that the participants rated the robot behavior as being slightly more important than its appearance for education (S24: $M = 4.84$, $SD = 1.59$), health/elderly care (S27: $M = 4.79$, $SD = 1.5$), drone-based applications (S26: $M = 4.24$, $SD = 1.71$), and entertainment (S25: $M = 5.04$, $SD = 1.76$).

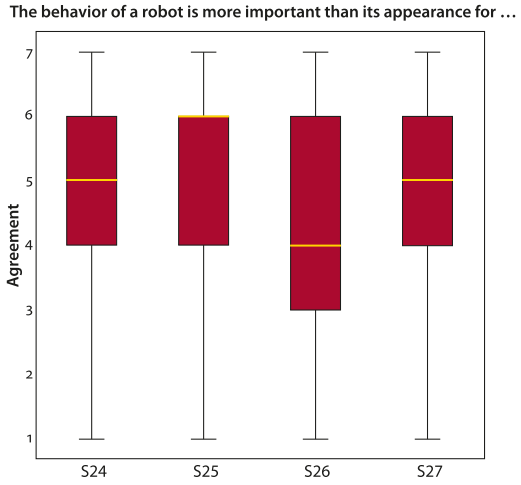


Figure 6. Medians and standard deviations for statements aiming to determine in which context the robot behavior is more important than its appearance. The complete statements are provided in Table (4)

Furthermore, to understand how social norms can be integrated into robots and how this affects human-robot interaction, the participants were asked to answer eight separate statements (S28-S35 in Table 4). Based on Figure 7, there is a slight agreement that the social norms that robots should follow are different from human social norms (S28: $M = 4.42$, $SD = 1.62$), and that a socially normative robot behavior cannot be hard-coded nor programmed but must be learned through interaction with humans (S29: $M = 4.82$, $SD = 1.6$). Moreover, the participants agreed that an appropriate robot behavior cannot, solely, be learned from watching human-human interaction but must be learned through interaction with humans (S30: $M = 5.48$, $SD = 1.29$), that social norms can be learned under conditions of uncertainty (S35: $M = 5.22$, $SD = 1.14$), that the evaluation of a robot behavior regarding its conformity to social norms must be done in natural human environments and cannot be done in artificial laboratory settings (S32: $M = 5.15$, $SD = 1.52$), and that allowing robots to take into account the descriptive characteristics of a user, e.g., appearance, to determine appropriate socially nor-

mative behaviors can lead to discrimination and racism (S33: $M = 5.3$, $SD = 1.41$). In contrast, the participants slightly disagreed that following social norms, i.e., what the group believes is appropriate, is more important than taking care of an individual's personal preferences (S31: $M = 3.7$, $SD = 1.49$) and that sharing experiences through cloud-based robot systems can lead to less personalized social behaviors (S34: $M = 3.82$, $SD = 1.56$).

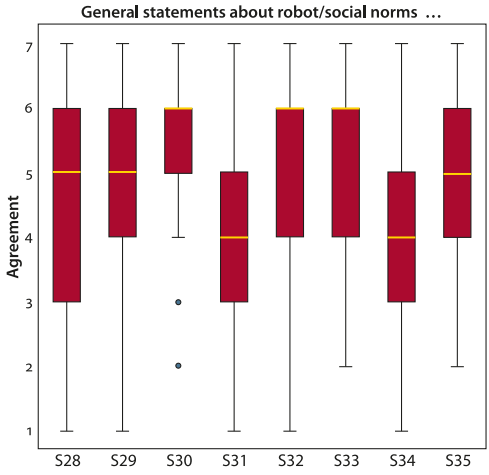


Figure 7. Medians and standard deviations for the eight separate statements about how social norms can and should be integrated into robots and how this affects human-robot interaction. The complete statements are provided in Table (4)

Benchmarks and metrics to evaluate socially normative robot behavior

There exist no specific benchmarks or metrics to objectively evaluate normative robot behavior and compare different approaches to integrate social norms into human-robot interaction. Thus, we asked the participants to rate 28 different statements regarding the characteristics that a good and general benchmark should have on a 7-point Likert scale ranging from "strongly disagree" to "strongly agree".

Figure 8 shows that according to the participants, the benchmark should include scenarios evaluating interaction in different social settings (S37: $M = 6.03$, $SD = 0.85$), interaction with different cultures and age groups (S38: $M = 5.93$, $SD = 0.88$), socially aware navigation (proxemics) (S41: $M = 6.18$, $SD = 0.94$), physical human-robot collaboration (S40: $M = 6.13$, $SD = 0.77$), and multi-party human-robot interaction (S39: $M = 5.92$, $SD = 0.95$). Overall, as is shown in Fig-

ure 9, the participants agreed that a good benchmark should consist of multiple scenarios that can be independently used (S36: $M = 5.75$, $SD = 1.02$), that the employed scenarios should allow active/online learning of normative behavior (S43: $M = 5.72$, $SD = 1.12$), and should be independent of a specific robot, like the RoboCup Open Platform League, to allow investigating the influence of the robot shape and design (S42: $M = 5.61$, $SD = 1.04$).

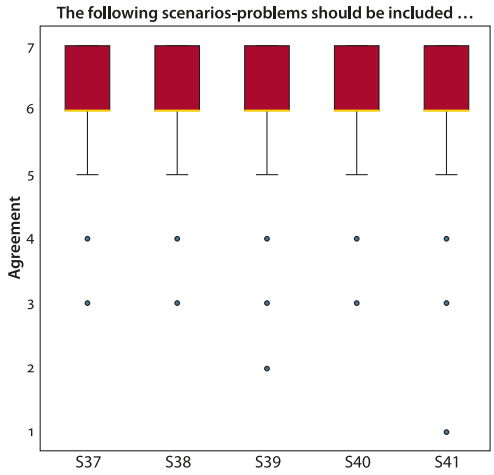


Figure 8. Medians and standard deviations for statements aiming to determine which scenarios should be included in a benchmark. The complete statements are provided in Table (5)

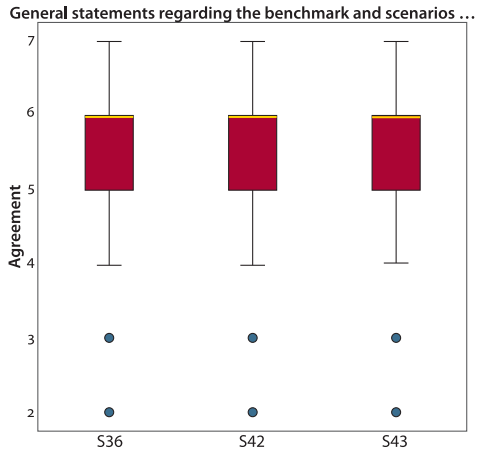


Figure 9. Medians and standard deviations for three separate statements about the characteristics a good benchmark and included scenarios should have. The complete statements are provided in Table (5)

Figure 10 shows that the participants agreed that robots should be evaluated regarding their adaptability to different environments (S46: $M = 5.93$, $SD = 1.08$) and users (S45: $M = 5.74$, $SD = 0.99$) as well as their safety (S44: $M = 5.86$, $SD = 0.98$), while the participants only slightly agreed that the benchmark should evaluate whether robots can handle multiple interactions simultaneously (S47: $M = 4.57$, $SD = 1.51$).

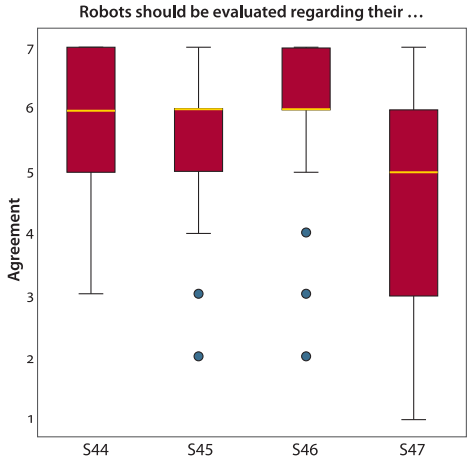


Figure 10. Medians and standard deviations for statements aiming to determine which criteria should be used to evaluate robots. The complete statements are provided in Table (5)

Based on the participants ratings in Figure 11, the most important criteria to evaluate human-robot interaction should be their similarity to human-human interaction (S48: $M = 6.38$, $SD = 0.81$). In addition, the participants agreed that it is important to evaluate interactions regarding their efficiency (S50: $M = 6.05$, $SD = 0.79$) and the amount of engagement of the user(s) (S49: $M = 5.99$, $SD = 0.85$), while the robot's success in achieving its goal is considered a less important criteria (S51: $M = 4.92$, $SD = 1.26$).

Moreover, the participants rated different parameters to determine which criteria should be used to evaluate how the robot is perceived. Figure 12 shows that the participants agreed that self-awareness (S52: $M = 6.19$, $SD = 0.96$), predictability (S54: $M = 6.02$, $SD = 0.9$), safety (S58: $M = 6.01$, $SD = 1.04$), trustworthiness (S57: $M = 5.68$, $SD = 1.22$), human-awareness (S53: $M = 5.64$, $SD = 0.94$), and human-likeness (S55: $M = 5.49$, $SD = 1.15$) should be considered when evaluating how the robot is perceived. Although, the participants agreed that flexibility (S59: $M = 4.96$, $SD = 1.53$) and friendliness (S56: $M = 4.29$, $SD = 1.62$) should be considered in the evaluation, these two parameters got slightly lower ratings.

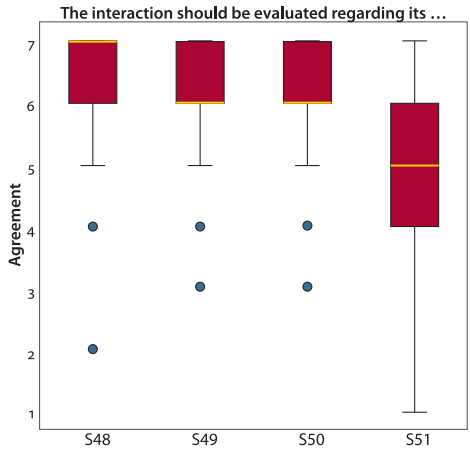


Figure 11. Medians and standard deviations for statements aiming to determine which criteria should be used to evaluate interactions. The complete statements are provided in Table (5)

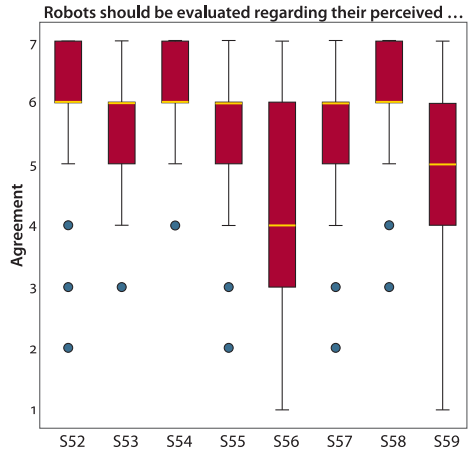


Figure 12. Medians and standard deviations for statements aiming to determine which criteria should be used to evaluate how the robot is perceived. The complete statements are provided in Table (5)

In addition, to determine which criteria should be used to evaluate how the social interaction is perceived, the participants rated different parameters. Figure 13 shows that the participants agreed that predictability ($M = 5.99$, $SD = 0.93$), efficiency ($M = 5.7$, $SD = 1.16$), naturalness ($M = 5.68$, $SD = 1.04$), and comfortability ($M = 5.6$, $SD = 1.28$) are all important criteria to evaluate how the social interaction is perceived.

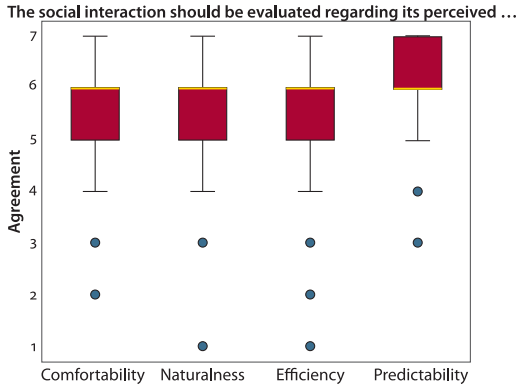


Figure 13. Medians and standard deviations for statements aiming to determine which criteria should be used to evaluate how the social interaction is perceived

Interactive session summary

For the interactive discussion, we asked the participants of the TSAR 2021 workshop[4] to split into two groups to discuss two important open-questions regarding the integration of social norms into human-robot interaction. The group discussions were limited to 20 minutes, after which we asked each group to present their results to the other group followed by the possibility for open discussion among all the workshop participants. The conclusions reached by both groups are summarized as follows.

Should we develop robot social norms or should robots follow human social norms?

The participants in the interactive group discussion reached the following conclusion regarding the above question. In general, we need to develop robot social norms that will be different from human social norms and will vary depending on the type and role of the robots, which is to some degree similar to differences in social norms for humans based on their role in society. For example, a quadrocopter delivering parcels will follow different social norms than a humanoid robot at the reception of a hotel. Only if a robot is indistinguishable from a human, will the social norms followed by the robot be the same as human social norms.

4. https://tsar2021.ai.vub.ac.be

Can socially normative robot behavior be hard-coded/programmed or must it be learned through interaction with humans?

The participants of the interactive discussion concluded regarding the above question that nowadays learning from scratch is not feasible in real environments, and that it is necessary to hard-code some basic norms that can then be adapted and optimized through learning. However, just learning, e.g., through reinforcement learning, is not enough, instead, robots should be able to reason about their own and human behaviors which would require proper grounding mechanisms to represent percepts through symbols and convert symbols to actuator commands. Furthermore, it is important that learning is continuous because norms can vary based on the situation and the people.

Panel discussion

In this section, we summarize the panel discussion with four renowned researchers held at the TSAR 2021 workshop in August 2021.

- **Prof. Kerstin Dautenhahn – University of Waterloo – Canada** is a research chair in Intelligent Robotics and director of Social and Intelligent Robotics Research Laboratory at University of Waterloo in Ontario, Canada, and a visiting professor at the University of Hertfordshire, UK. She received her master's and doctoral degrees from the University of Bielefeld, Germany. Her research focuses on discovering fundamental principles and mechanisms that can make robots more socially intelligent, as well as enabling them to interact with people in a trustworthy and efficient but also "natural" and socially acceptable manner. Her other interests include human-robot interaction, social robotics, assistive technology and artificial life.
- **Dr. Amit Kumar Pandey – beingAI – Hong Kong** is the co-founder and Chief Scientific Officer of beingAI, and founder and president of Socients AI and Robotics. He received his bachelor's degree from the Jaypee Institute of Information Technology, India, his master's degree from the Indian Institute of Technology, India, and his doctoral degree from LAAS-CNRS, France. His research focuses on addressing societal needs through scientific advancements, new technologies and ecosystem creation. His other interests include socially intelligent and socially interactive robots and AI agents, and the society.
- **Prof. Brian Scassellati – Yale University – USA** is a professor of computer science, cognitive science, and mechanical engineering at Yale University, USA, and Director of the NSF Expedition on Socially Assistive Robotics. He

received his bachelor's, master's, and doctoral degrees from the Massachusetts Institute of Technology, USA. His research focuses on building embodied computational models of human social behavior, evaluating models of how infants acquire social skills, and assisting in the diagnosis and quantification of disorders of social development, such as autism. His other interests include humanoid robots, human-robot interaction, artificial intelligence, machine perception, and social learning.

- **Prof. Greg Trafton – Naval Research Laboratory -USA** Greg Trafton is an affiliate/adjunct professor at George Mason University, USA, and the Section Head for the Intelligent Systems Section at the Navy Center For Applied Research in Artificial Intelligence in Washington, DC, USA. He received his bachelor's degree from Trinity University, San Antonio, TX, USA, and his master's and doctoral degree from Princeton University, USA. His research focuses on cognitive approaches to enable more intelligent systems that are able to work more effectively with people. His other interests include cognitive science, cognitive robotics and human-robot interaction, predicting and preventing procedural errors, the cognition of complex visualizations, interruptions and resumptions, and spatial cognition.

During the panel discussion, the panelists exchanged opinions with each other as well as other workshop attendees about several important topics regarding the employment of social norms in human-robot interaction. Following the main addressed points are summarized and the views expressed by the panelists highlighted.

1. **Do we need human social norms or robot social norms?**
   - **Dautenhahn:** Robots should have their own social norms because if they behave like humans, it might be confusing or even harmful. However, I would say that robots should be aware of human social norms because, at the end of the day, they need to fit in and if no one understands their behavior because it is so unlike what humans do, people might not find it socially acceptable. Therefore, a robot needs to be able to detect human social norms, make a judgment, and then based on its own social norms, it needs to be mapped to the expectations that people have with their human social norms. For example, if a group of children is playing together and slapping each other's back, a robot cannot just imitate this because it might be dangerous and one would not necessarily expect this kind of behavior from a robot.
   - **Pandey:** Social norms are not one set, there is no proper definition, no handbook of social norms. Thus, we need to determine which subset is rel-

evant for robots to understand and exhibit. If an artificial agent looks like a human but does not behave naturally, it will place some cognitive load on the user. For example, if the user is pointing at something or looking at some place and expecting the robot to do the same because people would do it.

- **Scassellati:** Maybe this is not the question we should be addressing. It comes down to this: a social norm is defined with respect to a particular group and what the question is really asking is: "Do you want these robots to be part of the group or not?" If you want them to be a part of the group, then they should follow the rules of that group because that is, in many ways, the definition of membership in a group. Why would we want them to be a part of a group? There are some benefits why we might want them to be in a group, e.g., we might want to treat them like peers in terms of education, attention, or ease of interaction, while there are reasons we might want them to be outsiders, e.g., we might want to order them around or turn them off without feeling bad. So, the question is, for whatever you want to do, is the robot a part of the group or not?
- **Trafton:** I think it is about how people perceive the robot. So, what if the robot is doing something, and people feel they need to follow along? I think, that if a robot is doing something that all robots are doing, then people will think should I do that too? Although it is easy to say no, I should not, some people would easily follow the robots, even if they don't look like people. Im worried that people will follow the robots a bit much.

2. **Should we see social norms as group norms, or should they be at an individual level, like personalized preferences? For example, if someone does not like to make eye contact, should the robot be designed not to make eye contact with them, even if most people prefer that the robot makes eye contact?**
   - **Dautenhahn:** I think one approach could be doing customization by human-robot teaching. People could teach the robot the norms that they would expect it to show, both on a group society level and individual level.
   - **Pandey:** From my perspective, the answer depends on the use case, i.e., what the robot is designed for. Should the robot facilitate a sort of social behavior in a person to do something according to a social norm? There is always a trade-off between personalization and how much robots should follow some standard rules they have been programmed for. For example, if a robot is supposed to care for someone with dementia, a professional healthcare provider will try to make eye contact to establish a sort of bond or trust to start communication. However, what if the person is trying to

make no eye contact? Should the robot give priority to the person or some high-level social norms? These are questions for which we don't have an answer, at least from my perspective.
- **Trafton:** That's what a norm is. A norm is a normal and not absolute behavior at some level. So, if you don't want the robot to make eye contact, it should not do it. It should be overridable to some degree.

3. **If we define robot-specific social norms, how do we ensure that there is an agreement across platforms? How do we ensure that there is not a robot social norm per manufacturer, for example?**
   - **Pandey:** Since there is no kind of handbook to handle all the defined social norms, I think it is more a question about the minimal set of safe behaviors, which are implementing a sort of social norm. To make sure that there is not a robot social norm per manufacturer, we need a global standard, otherwise, due to different robots and features, different behaviors will emerge. It is important, to be aware that the bias from the people behind the robot, e.g., programmers, might be reflected in the behavior of the robots.
   - **Scassellati:** I think we are unlikely to see any standardization. For example, it has been 40 years, and we have not been able to agree on what voltage to run robots at. Thus, the idea that we will all agree on how social norms should be designed, is unlikely to happen any time or in my lifetime.

4. **Is it beneficial to share experiences between robots, e.g., via the cloud, instead of having them learn separately?**
   - **Dautenhahn:** I think sharing experiences via the cloud is not feasible because it would require constant access to the internet, which is not always possible. Additionally, there is a huge privacy issue when sharing experiences and data from personal robots in schools or care facilities. However, if we assume that these problems don't exist, sharing knowledge, in general, is an interesting idea but it requires some deep thinking to avoid inconsistencies, when a system tries to combine different norms that should not be combined.
   - **Pandey:** This is a bigger question than, simply, social norm learning from a robot's context. What we are trying to think about is general artificial intelligence, which has all the knowledge through connected intelligence, and then robots will be simply a kind of body for that knowledge. There are, obviously, many issues to consider like privacy and security, but for particular practical applications, these kinds of robots might be useful.

For example, in the context of trans-cultural nursing where there are people from different cultural backgrounds, it might be useful if robots have shared knowledge about different cultural aspects and social norms to make the interaction more meaningful and grounded for a particular person.

- **Scassellati:** I think this comes down to let's first solve the problem of context and then this kind of cloud things will actually be useful because we will be able to say that it is a similar situation I can learn from but without context how do we know whether it is a good situation to learn from or not.
- **Trafton:** Different cultures have different social norms, and if you have robots across cultures, I think you will have a mess. I can see a group of robots in one geographical or social area, but I think that brings up the other question of do we really want to learn social norms.

5. **We have many social norms that we apply to other human beings, but what are social norms that humans have toward robots?**
   - **Dautenhahn:** Prof. Takayuki Kanda – from Kyoto University, Japan – often shows interesting videos about children abusing an innocent robot that is trying to interact with children in a shopping mall. Many children are very nice, but you also have children who start to hit the robot. What this illustrates is that you cannot predict what people will do when they meet a robot. Do they think it is an expensive device and they are not supposed to touch it because otherwise, their parents are getting angry, or just a toy they are allowed to squeeze, push, and pull? So, it is difficult to classify how people behave, and many people behave differently when they are alone than when they are in a group with other people. Even without any robot, we just change our behavior, and completely new behavior can emerge. It is made more complicated by the fact that we have such a huge design space of robots, e.g., humanoid robots, animal robots, and robots that people don't know what they are and have no reference or mental model for.
   - **Pandey:** I think it is too early to figure out and we are still not seeing the mass phenomenon of robots integrating with society. With time, we will see how we are integrating with robots and how robots are integrating with society, which might be similar to when smartphones came out, where we had first to understand it and adjust to it. Perhaps, once the novelty effect goes away, robots will be seen everywhere with different kinds of robots doing different things. So, perhaps the effect would be different. We want something which is simply useful, that's all. At the moment, peo-

ple are coming closer to a robot simply to touch or to take a selfie with it, but I am not sure this will be the same in the future. So, there is a long way to go to see the coexistence of humans and robots.

- **Scassellati:** I would guarantee that if in the US someone is going into a post office and sees a line of robots standing in a queue, they would just pass to go first, but I think you can get different results in different countries. I don't think anyone feels bad about that because everybody has a different justification for their behavior. For example, if someone believes everyone is equal and everyone's time is valuable, they would wait in a queue because that is the respectful thing to do, while if someone believes that for a robot, time does not matter but their time is important, they will just go to the front of the line. I think you will see differences in terms of what we do from a human perspective, I think these are very much certain things that people violate easily with robots.
- **Trafton:** I think normal people don't have expectations about robots because they all look different. We conducted a study about line-following robots. In one condition, we showed a video of a robot cutting the line, and people who watched the video got offended by the robot's behavior. While in the other condition, when the robot went to the end of the line, people just said, yeah, the robot got to the end of the line, what is the big deal? So, I think it matters whether we have strong expectations about the appropriate behavior in a specific situation.

6. **How can we evaluate robot behavior regarding its compliance to social norms, what kind of scenario should be covered, and should we try it in a real environment or the laboratory?**
   - **Dautenhahn:** There are different methodologies that are useful at different periods during system development, so there are often good reasons. One might choose simulations or online studies; however, if one wants to get real data, one needs to have people in the real environment; e.g., schools, hospitals, post offices, or hotels. This ideally occurs through observing them when they don't know they are being observed, which is difficult because one needs their approval. Even if one goes to the field, people know that it is a part of an experiment.
   - **Scassellati:** Novelty is a huge factor when doing experiments with social robots, and anytime one does something once, it is probable that the data is not good. We all run experiments in the laboratory and report them, but this is different from the way people will engage with social robots in their homes for a month. The very first time someone sees this brand new, cool, fancy, shiny robot will probably be significantly different. We are probably

going to spend the next 50 years undoing what were just novelty effects that we published during the last 20 years because we could not do longer studies at that time.

7. **Talking about the novelty effect, how can we ensure reproducibility, especially, when conducting experiments in real-world environments like private homes?**
   - **Dautenhahn:** I think everyone's home is different, e.g., different sizes, layouts, furniture, etc., but one could define the problem in a certain way depending on the task, how the robot should interact with the person, and what type of services it should provide. Certainly, it is easier nowadays because there are some robot platforms, such as Nao and Pepper, that many laboratories in the world have; however, I think that the main obstacle is that reproducibility studies will be difficult to publish because both journal and conference papers are all about novelty, which means that most graduate students will not be interested in this kind of studies. I think the point is that HRI is not about algorithms and how they perform, instead it is about the interaction, which cannot be fully controlled due to many emerging effects.
   - **Pandey:** I think this is where industrial research comes in, because we would like to replicate interesting results and try to do that in different environments once it is done well in the same environment.
   - **Scassellati:** I think we should not bother with reproducibility at this point. I don't think we know enough, and I don't think we have clearly established methods in this field, yet. So, at this point, we cannot worry about reproducibility, because, in some ways, we are still in the wild west exploring stage of this field. It is not just that we don't know what is the best thing the robot does, we don't know what the actual human social norms are supposed to be, and because of that, we are just not close to being standardized.
   - **Trafton:** I think HRI is still too young to be worried too much about replicability with all the other issues, but I don't want people to get this sense that we would all agree that replication is not important.

Discussion

The two main questions investigated through the survey, interactive session, and panel discussion are: (1) According to researchers in relevant communities, how does compliance to social norms influence the way humans perceive robots and

the corresponding interactions? and (2) what do researchers in relevant communities think are good benchmarks and metrics to support the evaluation of socially normative robot behavior?

Overall, the participants in the survey agreed that both robots that follow social norms and the corresponding interactions will be perceived more positively. For example, the robots will be perceived as more friendly, empathic, and trustworthy (Figure 2), while the interaction will be perceived as more enjoyful, natural, and comfortable (Figure 3) which together will lead to greater user satisfaction, longer interactions, and stronger human-robot relationships (Figure 4). The aim of enabling robots to follow social norms is to make robots behave more appropriately, however, what counts as appropriate depends, based on the responses to the survey, mainly on the robot's type and shape as well as its perceived gender, the human's personality and culture, and the relationship between the robot and the human, while it does only slightly depend on the gender of the human according to the participants (Figure 5). In general, the participants agreed that the behavior of robots is more important than their appearance, especially for entertainment (Figure 6).

Another important question is whether robots should follow the same social norms humans follow or whether they should have their own social norms, i.e., are there situations for which the appropriate behavior is different for robots and humans? Most participants in the survey slightly agreed that robots should follow different social norms; however, there were a number of participants who slightly disagreed with this (Figure 7). The participants of the interactive session and the panelists mostly agreed that robots should have their own social norms, which will vary depending on the type and role of robots, similar to humans who are following different social norms in different societies or depending on their roles in society. However, robots should only have their own social norms as long as they are clearly distinguishable from humans, when this is no longer the case, they should follow the same social norms as humans. Important to note is that robots still need to understand human social norms, even if they are not following them, to better understand and predict human behavior.

Both the participants of the interactive session and the survey agreed that socially normative behavior should be learned, however, since learning from scratch is very difficult, some initial behaviors must be hard-coded and then autonomously and continuously optimized through learning to be able to incorporate information from new situations. Most of the survey participants believed that social norms can be learned under conditions of uncertainty and that learning needs to be done during interaction with humans because learning social norms by just watching human-human interaction is not possible (Figure 7). This is consistent with the idea that robots should develop their own social norms,

which can certainly not be learned from observation of human-human interaction but requires robots to interact with humans and observe the feedback provided during and after the interaction to adjust and optimize their social behavior.

When enabling robots to learn social behavior, it is important to consider whether they should optimize their behavior toward social norm compliance or whether they should try to optimize their behavior based on the preferences of the individuals they are interacting with. The majority of the participants in the survey neither preferred robots to give priority to social norms nor personal preferences, although there was a slight preference toward personal preferences, which is consistent with the arguments provided during the panel discussion that the answer depends strongly on the specific situation the robot is in but that personal preferences should usually override default normative behavior. There was strong skepticism among the panelists about the feasibility to share experiences between robots, especially when they are deployed in different societies or cultures, which is consistent with the strong uncertainty shown by the survey participants regarding whether sharing experiences will lead to less personalized robot behaviors, although there was a slight tendency to believe that this would not be the case. However, independent of the potential technical difficulties, the panelists pointed out that the potential privacy implications, especially when considering robots in private homes or care facilities, and challenges due to limited internet connectivity, if sharing of experiences requires constant internet access, should not be underestimated.

When thinking about the evaluation of different robot behavior regarding its conformity to social norms, an important question is whether such an evaluation can be performed in artificial laboratories or in simulation or whether it will require robots to be deployed in natural human environments, e.g., schools, hospital, care facilities, or shopping malls. The majority of the participants of the survey believed that conformity to social norms must be evaluated in natural human environments, which is in general consistent with the views of the panelists, although they mentioned that evaluations in the lab can be useful during certain phases of system development and that natural environments do not necessarily lead to completely unbiased results if the people interacting with the robot are aware that the interaction is part of an experiment.

Independent of where experiments and evaluations are conducted, it is important to ensure that experiments done by different researchers and in different parts of the world can be compared to simplify the identification of methods and models that work well across different scenarios, thereby, fostering progress in the field. One possibility to achieve this is to define general benchmarks and metrics. Currently, no such benchmarks exist and according to the panelists, it might still be too early for the field to care about reproducibility. Nevertheless,

it is beneficial for the field to understand what characteristics a good benchmark should have and what criteria should be used to evaluate socially normative robot behavior so that this information is available when the community decides that the field is ready for it. Therefore, one goal of the survey was to gather the views in related research communities about the characteristics such a benchmark should have. The participants in this survey agreed that a good general benchmark should have multiple scenarios that can be independently used, that the scenarios should be independent of a specific robot to allow investigating the robot's shape and design, and that the scenarios should allow active or online learning of normative behavior (Figure 9). They agreed that the benchmark should cover (at least) the following scenarios/problems: different social settings, like teacher-student, guide, social companion, etc., different cultures and age groups, multi-party human-robot interaction, physical human-robot collaboration, and socially aware navigation (proxemics) (Figure 8).

According to the survey participants, robots should be evaluated regarding their safety as well as their adaptability to different users and environments, while evaluating robots whether they can handle multiple interactions simultaneously was seen as not as important (Figure 10). This highlights the importance of learning socially normative behavior to achieve the required adaptability to different users and environments. Furthermore, the survey participants agreed that robots should be evaluated regarding their perceived self-awareness, human-awareness, predictability, human-likeness, trustworthiness, and safety, while evaluating them regarding their perceived friendliness and flexibility was considered not as important (Figure 12). Especially, the human-likeness criteria is interesting because this would mean that robots that look and behave more like humans would be rated better, although based on the survey, interactive session, and panel discussion the aim should be to determine robot social norms that ensure that robots behave in a way that is considered appropriate by humans, which most likely will be different in some situations from the behavior expected of humans. This shows that the criteria determined through the survey are only a starting point and that further discussions within the community are necessary to determine the best set of evaluation criteria.

In addition to the robot, it is important to evaluate the interaction, and based on the survey responses, the most important criteria is the similarity to human-human interaction, which raises the question of whether the aim is to let robots learn their own social norms, or whether they should "just" try to follow human social norms. Further criteria that the participants considered important to evaluate human-robot interaction are the efficiency of the interaction, user engagement, and whether the robot achieves its goal, although the latter was rated as less important than the other criteria (Figure 11). Finally, the survey participants rated

different criteria to evaluate how the social interaction is perceived. The obtained ratings showed that social interactions should be evaluated regarding their perceived predictability, efficiency, naturalness, and comfortability (Figure 13).

Overall the results of the interactive discussion, panel discussion, and survey show that enabling robots to follow social norms has the potential to strongly improve human-robot interactions and that there is a strong need to start thinking and talking about how to best evaluate socially normative behavior because it is non-trivial and will require a collective effort by the whole (or at least large parts) of the community. It is important to keep in mind that the survey only provides an overview of the views of part of the community and is not necessarily representative of the community as a whole. Furthermore, the survey does not directly answer how compliance to social norms changes the way humans perceive robots and the corresponding interactions but only indicates what changes the community expects; therefore, there is a need to verify these hypotheses through carefully designed experiments to ensure that future research is not based on wrong assumptions.

Conclusions

This paper presented and discussed the views of researchers in relevant research communities regarding the effects of socially aware robot behavior on the perception of robots and corresponding interactions. Furthermore, this study provided an overview of which characteristics and metrics the wider research community believes constitute a good general benchmark for the objective evaluation of socially aware robot behavior. Finally, based on the presented findings we provided some suggestions for future research toward socially aware robot behavior. In future work, we are planning to work on verifying the hypotheses the research community believes to be true through carefully designed experiments and to work on the development of a general benchmark for the evaluation of socially normative robot behavior together with the community.

Acknowledgement

The authors would like to thank the panelists, the workshop participants, and the survey participants for their participation because this work would not have been possible without them.

References

Cabin, R. J., & Mitchell, R. J. (2000). To Bonferroni or not to Bonferroni: when and how are the questions. *Bulletin of the ecological society of America*, *81*(3), 246–248.

Ciou, P.-H., Hsiao, Y.-T., Wu, Z.-Z., Tseng, S.-H., & Fu, L.-C. (2018). Composite reinforcement learning for social robot navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (p. 2553–2558). https://doi.org/10.1109/IROS.2018.8593410

Dautenhahn, K., Walters, M., Woods, S., Koay, K. L., Nehaniv, C. L., Sisbot, A., … Simeon, T. (2006). How may I serve you? A robot companion approaching a seated person in a helping context. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (p. 172–179). https://doi.org/10.1145/1121241.1121272

Gao, Y., Yang, F., Frisk, M., Hemandez, D., Peters, C., & Castellano, G. (2019). Learning socially appropriate robot approaching behavior toward groups using deep reinforcement learning. In *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (p. 1–8). https://doi.org/10.1109/RO-MAN46459.2019.8956444

Joosse, M., Lohse, M., & Evers, V. (2014). Lost in proxemics: spatial behavior for cross-cultural HRI. In *Proceedings of the HRI 2014 Workshop on Culture-Aware Robotics* (p. 1–6).

Koay, K. L., Syrdal, D. S., Walters, M. L., & Dautenhahn, K. (2007). Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (p. 564–569).

Kruse, T., Pandey, A., Alami, R., & Kirsch, A. (2013, December). Human-Aware Robot Navigation: A Survey. *Robotics and Autonomous Systems*, *61*(12), 1726–1743. https://doi.org/10.1016/j.robot.2013.05.007

Li, D., Rau, P.-L., & Li, Y. (2010). A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, *2*(2), 175–186. https://doi.org/10.1007/s12369-010-0056-9

Mensah, Y. M., & Chen, H.-Y. (2013, April). Global Clustering of Countries by Culture – An Extension of the GLOBE Study. *SSRN*.

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral ecology*, *15*(6), 1044–1045. https://doi.org/10.1093/beheco/arh107

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, *316*(7139), 1236–1238. https://doi.org/10.1136/bmj.316.7139.1236

Sarathy, V., Wilson, J. R., Arnold, T., & Scheutz, M. (2016). Enabling basic normative HRI in a cognitive robotic architecture. *arXiv preprint arXiv:1602.03814*.

Takayama, L., & Pantofaru, C. (2009). Influences on proxemic behaviors in human-robot interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (p. 5495–5502). https://doi.org/10.1109/IROS.2009.5354145

Tomic, S., Pecora, F., & Saffiotti, A. (2014, August). Too Cool for School – Adding Social Constraints inHuman Aware Planning. In *The 9th International Workshop on Cognitive Robotics (CogRob)*. Prague, Czech Republic.

Trovato, G., Zecca, M., Sessa, S., Jamone, L., Ham, J., Hashimoto, K., & Takanishi, A. (2013). Cross-cultural study on human-robot greeting interaction: acceptance and discomfort by Egyptians and Japanese. *Paladyn, Journal of Behavioral Robotics*, *4*(2), 83–93. https://doi.org/10.2478/pjbr-2013-0006

Wang, L., Rau, P.-L. P., Evers, V., Robinson, B. K., & Hinds, P. (2010). When in Rome: the role of culture & context in adherence to robot recommendations. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (p. 359–366). https://doi.org/10.1145/1734454.1734578