What a Pity, Pepper!

How Warmth in Robots' Language Impacts Reactions to Errors during a Collaborative Task

Laura Hoffmann
Bielefeld University, Germany
lahoffmann@techfak.uni-bielefeld.de

Melanie Derksen
Bielefeld University, Germany
melanie.derksen@uni-bielefeld.de

Stefan Kopp
Bielefeld University, Germany
skopp@techfak.uni-bielefeld.de

ABSTRACT

We investigate the impact of warmth in robots' language on the perception of errors in a shopping assistance task (*N*=81) and found that error-free behavior was favored over erroneous if the dialogue is machine-like, while errors do not negatively impact liking, trust and acceptance if the robot uses human-like language. Warmth in robots' language thus seems to mitigate negative consequences and should be considered as a crucial design aspect.

KEYWORDS

human-robot interaction; errors; warmth; trust; social cognition

ACM Reference format:

Laura Hoffmann, Melanie Derksen and Stefan Kopp. 2020. What a Pity, Pepper! How Warmth in Robots' Language impacts Reactions to Errors during a Collaborative Task. In *HRI '20 Companion: ACM/IEEE International Conference on Human-Robot Interaction, March 23–26, 2020, Cambridge, UK. ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/3371382.3378242

1 Motivation and Hypotheses

In real world HRI, errors are likely to occur and should be taken into account when designing interactions as they affect the development of trust [1,2], which is a necessary prerequisite for acceptance [3]. Interestingly, previous work revealed that errors negatively affect peoples' performance and trust in a robot [4,5,6,7,8], whereas liking of the robot was often higher after failure than after flawless performance [4,6,9]. Attempts to explain these non-intuitive findings refer to the Pratfall effect, i.e., failure boosts liking [9]. Thinking of robots that should assist humans in the real world, e.g., remind them to take their medication, it is difficult to accept the notion that people favor erroneous over flawless robots. Another possible explanation for the observed mild evaluations could be the perceived warmth of a robot. The impact of warmth and competence has been investigated in the realm of the *Stereotype Content Model* (SCM: [10]), which assumes that specific combinations of warmth and competence determine emotional reactions, e.g., a person that is regarded as cold and incompetent evokes contempt, whereas a person that is perceived as incompetent but warm evokes pity. The SCM has already been studied in HRI [11]. Applied to robots' evaluations, the model suggests that erroneous robots evoke pity as long as they are perceived as friendly. However, robots' warmth has not yet been researched in the realm of robots' errors. We were thus curious to test whether the evaluation of a failing robot depends on its warmth. We conducted an experimental study that altered both factors systematically. We hypothesized that *pity (H1), likability (H2), trust (H3), and acceptance (H4) vary as a function of warmth and competence*, namely, erroneous behavior only elicits favorable reactions under warm conditions.

HRI '20 Companion, March 23–26, 2020, Cambridge, United Kingdom

ACM ISBN 978-1-4503-7057-8/20/03. https://doi.org/10.1145/3371382.3378242

2 Method and procedure

We conducted a laboratory experiment with a 2 (manipulated competence, MC: error-free/erroneous) x 2 (manipulated language warmth, MLW: machine-like/human-like [12]) between-subjects design, in which participants went grocery shopping with the assistance of a humanoid robot (Softbank's Pepper).

After a short briefing, subjects signed informed consent before they were guided to the robot, which introduced itself and the task. When the task was clear, the experimenter left to enable an undisturbed experience. The robot then moved along a predefined path and told the subject which products they needed. For each item the robot stopped and requested it, either using machine-like *("Item number 3: bar of chocolate. Please identify product...")* or human-like *("Yummy, the next thing we need is a bar of chocolate. I wonder who added this to the list...")* language. The interaction continued until eight items were in the robot's basket. In the error-free condition, the robot then asked to go to the cashier. In the erroneous condition, the robot dropped the basket and asked to repeat (machine-like: *"Error 1.0.5, basket missing. To repeat say: reboot..."*; human-like: *"Oh no, I lost the basked. Can we try it again?"*). Regardless of the subject's reaction, the experimenter terminated the interaction after the first trial and asked to complete the survey before the subjects were debriefed and received monetary compensation.

Dependent variables were collected in a survey using 5-point scales. Items of subscales were collapsed into scores. To test the effectiveness of our manipulation, we measure *perceived warmth* (7 items, e.g., cold/empathetic, α = .84) and *perceived competence* of the robot (7 items, e.g., incompetent/competent, α = .85) with a

Table 1. Descriptive statistics and ANOVA results for all dependent measures

	IV: Manipulated Competence (MC)	IV: Manipulated language warmth (MLW)						ANOVA (main and interaction effects)			
		machinelike		humanlike		MLW total					
Dependent variable		*M*	*SD*	*M*	*SD*	*M*	*SD*		$F_{(1,77)}$	*p*	η_p^2
Perceived competence	error-free	4.07	0.71	4.16	0.52	4.11	0.62	MLW	5.00	<.05	.06
	erroneous	3.54	0.74	4.12	0.66	3.84	0.75	MC	3.75	.06	.05
	MC total	3.81	0.76	4.14	0.59			MLW x MC	-	-	-
Perceived warmth	error-free	2.89	0.63	3.52	0.52	3.20	0.66	MLW	38.58	<.001	.33
	erroneous	2.79	0.56	3.66	0.44	3.24	0.67	MC	-	-	-
	MC total	2.84	0.59	3.59	0.48			MLW x MC	-	-	-
Pity	error-free	1.40	0.82	1.20	0.52	1.30	0.69	MLW	-	-	-
	erroneous	1.80	1.24	2.00	1.27	1.90	1.24	MC	7.08	<.01	.08
	MC total	1.60	1.06	1.61	1.05			MLW x MC	-	-	-
Likability	error-free	4.38	0.64	4.62	0.48	4.50	0.57	MLW	16.40	<.001	.18
	erroneous	3.93	0.61	4.67	0.42	4.31	0.64	MC	-	-	-
	MC total	4.16	0.66	4.64	0.45			MLW x MC	4.24	<.05	.05
Affect-based trust	error-free	2.84	0.87	2.95	0.69	2.90	0.78	MLW	10.07	<.01	.12
	erroneous	2.30	0.63	3.16	0.52	2.74	0.71	MC	-	-	-
	MC total	2.57	0.80	3.06	0.61			MLW x MC	6.06	<.05	.07
Cognition-based trust	error-free	4.44	0.54	4.09	0.55	4.27	0.56	MLW	-	-	-
	erroneous	3.46	0.51	3.78	0.47	3.62	0.51	MC	32.08	<.001	.29
	MC total	3.95	0.72	3.93	0.53			MLW x MC	8.54	<.01	.10
Intention to use	error-free	3.65	1.14	3.35	1.27	3.50	1.20	MLW	4.46	<.05	.06
	erroneous	2.35	1.04	3.67	0.86	3.02	1.15	MC	4.17	<.05	.05
	MC total	3.00	1.26	3.51	1.08			MLW x MC	11.28	<.01	.13

self-constructed semantic differential based on [10,13]. *Pity* was assessed with a single item. *Liking* was assessed with the likeability subscale from [14] (5 items, e.g., dislike/like; α = .87). Further, items from [15] were used to measure *cognition-based* (6 items; α = .76) and *affect-based trust* (6 items; α = .76). *Acceptance* was determined by subjects' compliance to the robot's request to restart after error (yes/no; error condition only), and their self-reported intention to use the robot in the future (single item). Subjects were N = 81 students (33 male) between 18 and 35 years (M = 25.95, SD = 3.87). Assignment to the conditions was random, and the experimenter was kept blind to the conditions.

3 Results and Discussion

To test our hypotheses, two-factorial ANOVAs with MC and MLW as fixed factors were calculated. Regarding the success of our manipulations, ANOVA revealed that *perceived warmth* was higher for human-like compared to machine-like language (Table 1). The *perceived competence* was (marginally) affected by MC and also MLW: the error-free robot was rated more competent than the erroneous. However, the erroneous robot was still rated highly competent, which might be due the successful transportation of items that the subjects experienced before the error (Table 1). Higher *perceived competence* ratings for the human-like compared to the machine-like robot further show that human-like language not only increased perceived warmth but also perceived competence. Regarding *pity*, no interaction effect emerged (H1), but a main effect of MC: The erroneous robot evoked more *pity* than the error-free (Table 1), while the mean ratings show overall low feelings of pity in all conditions. However, in line with [10], most pity was evoked by the warm and erroneous robot. Concerning *likability* (H2), we observed a main effect of MLW, indicating that a human-like speaking robot was more likable than a machine-like one (Table 1). Further, a significant interaction effect followed: In the machine-like conditions, the error-free robot was more liked than the erroneous (p<.001), but in the human-like conditions no significant difference emerged. For *cognition-based trust* (H3a), a significant main effect of MC emerged: The error-free robot was rated more reliable than the erroneous (Table 1). Again, a significant interaction effect appeared, showing that in the machine-like conditions, subjects trusted the error-free robot significantly more than the erroneous (p<.001), while in the human-like conditions, no significant difference emerged. The same pattern was observable for *affect-based trust* (H3b). Besides a significant interaction effect, a main effect of MLW occurred: More faith and attachment were attributed to the humanlike than the machine-like robot (Table 1). With respect to acceptance, almost all (38/41) agreed to repeat the assistance (H4a) when the robot dropped the basket. Only three refused to repeat the interaction, however, it is noteworthy that all of them where in the cold condition [Pearson's χ^2 (1,41) = 3.40, p = .065]. Regarding participants' *intention to use the robot again* (H4b), we observed significant main effects of MLW and MC (Table 1). Subjects were more willing to use the human-like and error-free than the machine-like and erroneous robot. A significant interaction effect demonstrated that subjects in the machine-like condition indicated a higher intention to use an error-free robot compared to an erroneous (p<.01), whereas in the human-like condition no such difference appeared.

In summary, we did not find higher sympathy for an erroneous robot as reported previously [4,6,9]. Furthermore, we did not observe significantly higher pity towards an erroneous robot that uses human-like language as hypothesized. Our findings, however, revealed that negative consequences of robots' errors on liking, trust and acceptance can be compensated by using human-like language. If the robot used human-like language, the impact of the error disappeared, indicating that such a socio-communicative capability is able to compensate for physical incapabilities in the case of the humanoid robot Pepper. Our findings further demonstrated that machine-like language can be helpful to reduce 'overtrust' in robots (e.g., as in [8]), because participants trusted the erroneous and machine-like robot less than the error-free one. Future studies should test whether the compensatory power of human-like language applies to other robots, and if it is useful to communicate a robot's limitations to users in erroneous situations.

REFERENCES

[1] Donald A. Norman. 2002. The Design of Everyday Things. Basic Books, Inc., New York, NY, USA.

[2] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[3] Kewen Wu, Yuxiang Zhao, Qinghua Zhu, Xiaojie Tan, and Hua Zheng. 2011. A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. *International Journal of Information Management* 31, 6 (2011), 572–581.

[4] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 13–323.

[5] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction,* ACM, 141–148. DOI: https://doi.org/10.1145/2696454.2696497

[6] Marco Ragni, Andrey Rudenko, Barbara Kuhnert, and Kai O Arras. 2016. Errare humanum est: Erroneous robots in human-robot interaction. In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2016)*. IEEE, 501–506. DOI:10.1109/ROMAN.2016.7745164

[7] Rik van den Brule, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, and Pim Haselager. 2014. Do robot performance and behavioral style affect human trust? *International journal of social robotics* 6, 4 (2014), 519–531.

[8] Paul Robinette, Ayanna Howard, and Alan R Wagner. 2017. Conceptualizing overtrust in robots: Why do people trust a robot that previously failed? In *Autonomy and Artificial Intelligence: A Threat or Savior?* Springer, 129–155.

[9] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4, 21 (2017). DOI: 10.3389/frobt.2017.00021

[10] Susan T Fiske. 2018. Stereotype content: Warmth and competence endure.Current directions in psychological science 27, 2 (2018), 67–73. DOI: https://doi.org/10.1177/0963721417738825

[11] Raquel Oliveira, Patrícia Arriaga, Filipa Correia, and Ana Paiva. 2019. The Stereotype Content Model Applied to Human-Robot Interactions in Groups. In *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI'19)*. IEEE, 123–132.

[12] Aike C Horstmann, Nikolai Bock, Eva Linhuber, Jessica M Szczuka, Carolin Straßmann, and Nicole C Krämer. 2018. Do a robot's social skills and its objection discourage interactants from switching the robot off ? *PloS one* 13, 7 (2018), e0201581. DOI:https://doi.org/10.1371/journal.pone.0201581

[13] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The robotic social attributes scale (rosas): Development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. ACM, 254–262.

[14] 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.

[15] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53, 6–8.