Signaling Robot Trustworthiness: Effects of Behavioral Cues as Warnings

Rik van den Brule[1,2], Gijsbert Bijlstra[2], Ron Dotsch[2], Daniël H.J. Wigboldus[2], and Pim Haselager[1]

[1] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, The Netherlands
r.vandenbrule@donders.ru.nl
[2] Behavioural Science Institute, Radboud University Nijmegen, The Netherlands

Abstract. By making use of existing social behavioral cues, domestic robots can express their uncertainty about their actions. Ideally, when a robot can warn that a mistake is likely, humans can take preemptive actions to ensure the successful completion of the robot's task. In the present research we show that a robot employing behavioral cues predictive of making mistakes is judged as more reliable and understandable than robots that do not use these cues predictively, and is trusted on par with a robot that does not employ behavioral cues.

1 Introduction

Domestic robots will face tasks in widely varying circumstances, ranging from an empty house to, for instance, a children's party. In certain situations a robot's functioning may require additional attention from those present. Minor interventions by humans might enable robots to carry out their task successfully. Therefore, a domestic robot may benefit from sending out warning signals when unexpected circumstances arise or a mistake becomes likely.

Current domestic robots already employ warning signals. However, an important drawback is that such signals are mostly used *after* a malfunction has occurred. Ideally, a robot would send online, proactive communication about uncertainty levels so that its owner can take preemptive action. Although such mistake-detection systems do not exist presently, it is possible to explore the feasibility of such systems by manipulating a robot's warning signals and the likelihood of mistakes following them in an experimental setting.

Robots can communicate uncertainty about their actions in many different ways. In this project, in which we borrow ideas from social psychology, we use nonverbal behavioral cues that occur naturally in human interaction. The advantage of employing existing behavioral cues to express uncertainty is that they are easily interpretable. Such attributes can influence a robots trustworthiness [1,2].

By utilizing uncertainty cues as a warning that a mistake is more likely to occur, the perceived reliability and understandability of a robot's can be increased.

G. Herrmann et al. (Eds.): ICSR 2013, LNAI 8239, pp. 583–584, 2013.

The robot's trustworthiness can be regulated by making the robot appear to be less trustworthy at the right times. In other words, the robots trustworthiness can be calibrated such that a users current level of trust is well mapped to the system's capabilities in the current circumstances [3].

2 Current Research

We tested whether the predictability of a robot's task performance with behavioral cues expressing uncertainty affects trustworthiness judgments in Human-Robot Interaction by means of a video study with a simulated Nao robot. The robot selected one of three soda cans by means of gaze behavior, after which it did or did not display a cue (the robot wipes its forehead with its hand). Participants then predicted whether the robot would point towards the selected can, or (accidently) push it off the table, after which they observed the robots action. After completing 60 of such trials, the participants rated the robot on a trustworthiness scale, as well as perceived reliability and perceived understandability scales based on [4].

82 participants were assigned to one of four conditions: predictive cues, no cues, random cues, or always cues. Importantly, only in the predictive condition the probability of a mistake given a behavioral cue was above chance level.

Analyses of variance (ANOVA) indicated that there was a significant effect of the presentation of the cue on trustworthiness, perceived reliability and perceived understandability, all $F(3, 78)$s > 5.96 , all ps $< .0010$. Post hoc Tukey's HSD tests revealed that participants in the predictive cues condition perceived the robot as more understandable and reliable compared to all other conditions, although for reliability the difference between no cues and predictive cues was marginally significant. There was a somewhat different pattern for trustworthiness. In the predictive condition, the robot was rated as significantly more trustworthy compared to the non-predictive conditions in which the behavioral cue was shown, but ratings did *not* differ significantly from the no cues condition. This suggests that a robot's trustworthiness is not based on a robot's behavior or predictability alone, but that the relationship between these attributes is more complex. In conclusion, signaling trustworthiness with behavioral warning cues is a promising, yet challenging, way to calibrate trust in Human Robot Interaction.

References

1. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., de Visser, E.J., Parasuraman, R.: A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. Human Factors 53(5), 517–527 (2011)
2. Van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Haselager, W.F.G.: Human-robot trust: A multi-method approach (submitted)
3. Lee, J., See, K.: Trust in Automation: Designing for Appropriate Reliance. Human Factors 46(1), 50–80 (2004)
4. Madsen, M., Gregor, S.: Measuring Human-Computer Trust. In: Proceedings of the 11th Australasian Conference on Information Systems, pp. 6–8 (2000)