Abstract

Whether a concept of "self" is necessary for understanding human behavior is still debated. Nonetheless, it has significance in everyday life: Lay individuals ascribe selves to humans, nonhuman animals and technical systems alike, shaping their interaction accordingly. While a layperson may not be able to provide a clear definition of a "self", they most likely use a naïve concept of selfhood to explain their own and others' behavior. The literature suggests that there are objective behavioral cues that elicit this attribution of selfhood, which may be as minimal as simple movement perceived as causal to some event. The present thesis aimed to identify which types of behavioral cues may increase selfhood-attributions to other agents such as robots. Specifically, this was done by comparing behavior of non-humanoid robots suggesting either the presence or absence of behavioral cues for one of the potential core characteristics identified in the literature.

Study 1 investigated the characteristics of causal behavior, equifinality, behavioral efficiency, learning sensitivity and context sensitivity. Results showed a consistent pattern of increased selfhood-attribution towards robots exhibiting any one of the examined minimal characteristics. Furthermore, most of the perceived behavioral characteristics of the robot were triggered by any single characteristic's cue. These results reflect a pattern similar to what has been referred to as the Halo effect: Even a single cue of selfhood related characteristics may be sufficient to trigger a change in overall selfhood-attribution to robots. This is framed in a Brunswikian model of selfhood-judgement, wherein selfhood is attributed based on the activation of self-related characteristics. However, the model reflects that not all of the perceived characteristics are directly triggered by their corresponding behavioral cues, rather that the characteristics interact with each other. The alternative possibility is also discussed that the concepts of these characteristics overlap more than is reflected in their use in language.

Either way, this study shows that people go way beyond the information given when attributing selfhood.

While the previous study only investigated non-social characteristics, the importance of social interaction for the emergence and functions of the "self" has been stressed in the literature. Thus, Study 2 investigated potential core characteristics of a social nature. While the general design was the same as in the previous study, the robots in this study were manipulated to suggest the presence or absence of the characteristics of social sensitivity, attention sharing and helping behavior. The results replicated the findings of the previous study that the presence of a single cue is sufficient to over-generalize to other, non-manipulated cues (framed as a Pars-Pro-Toto account of selfhood-attribution). Thus, Study 2 shows that this account can also explain the process of selfhood-attribution in social conditions. Interestingly, however, the over-generalization was stronger within the social domain, while the presence of other agents reduced the attribution of agency. This is discussed as reflecting how sociality might be construed vis-à-vis individual goal pursuit.

Considering that cueing a small aspect of selfhood is sufficient to trigger the entire selfhood concept with all its implications, the predication was made that single selfhood cues are as efficient as multiple selfhood cues in eliciting selfhood-attributions –as opposed by a cumulative attribution, wherein more cues would elicit stronger selfhood-attribution. Thus, study 3 compared in three experiments ratings for a robot exhibiting behavioral cues of efficient behavior, learning sensitivity and equifinality with ratings for a robot exhibiting only one of these cues. Contrary to the prediction, participants did not show the same degree of selfhood-attribution to both robots, but overall stronger over-generalization of self-related characteristics along with stronger attributions of selfhood towards the single-cue robot. This suggests that the Pars-Pro-Toto account of selfhood-attribution is missing the complete picture in the description of merely an over-generalization based on a single cue. The preference to over-generalize

agents showing cues for only a single characteristic is discussed as behavior that is potentially identified as more human-like as proposed by the literature on cognitive bootstrapping, which suggests that humans prefer to reuse knowledge or skills to form complex ideas and actions. Further, this study concludes that selfhood functions primarily as a social concept used for explaining human behavior and that future research should focus on this aspect rather than trying to identify a single scientific definition.