# Lay Causal Explanations of Human vs. Humanoid Behavior

Sam Thellman ✉, Annika Silvervarg & Tom Ziemke

Department of Computer and Information Science, Linköping University, Linköping
{sam.thellman, annika.silvervarg, tom.ziemke}@liu.se

**Abstract.** The present study used a questionnaire-based method for investigating people's interpretations of behavior exhibited by a person and a humanoid robot, respectively. Participants were given images and verbal descriptions of different behaviors and were asked to judge the plausibility of seven causal explanation types. Results indicate that human and robot behavior are explained similarly, but with some significant differences, and with less agreement in the robot case.

**Keywords:** Human-robot interaction, attribution, behavior explanation

## 1 Introduction

There is ample evidence that people interpret and explain the behavior of robots and other artificial agents using common-sense folk-psychological concepts such as beliefs, desires, intentions, and emotions [1, 2, 3]. Consider a robot bumping into a person. Such behavior could be explained as caused by a *goal* to bump into people, by a *temporary state* of sensory confusion, by a *behavioral disposition* from bad design, or by an *event* outside the robot's control (e.g. a slippery floor). Mental state attributions fundamentally shape people's interaction with others in that they set the course for how people perceive and respond to behavior. So far, however, there has been very little comparative research on how people actually interpret the behavior of different types of artificial agents, and how this compares to human-human social interaction. This study is intended to make a small contribution towards closing that gap by investigating people's lay causal explanations of human vs. humanoid behavior.

Fritz Heider, the founder of attribution theory, suggested that people have internalized and mastered a causal network of formal connections between concepts which underlies their social understanding of the world [4]. Böhm and Pfister recently proposed and empirically validated a model of people's lay causal explanations of human behavior, the *causal explanation network* (CEN) model (cf. Fig. 1), which suggests that people's explanations follow a specific inference pattern [5]. The model is based on previous attribution research and specifies seven cognitive categories that are assumed to be used for both behavior encoding and explanation: *goals*, *intentional actions*, *action outcomes*, *temporary states*, *dispositions*, *uncontrollable events*, and *stimulus attributes*. The categories are related to each other through "inference rules"

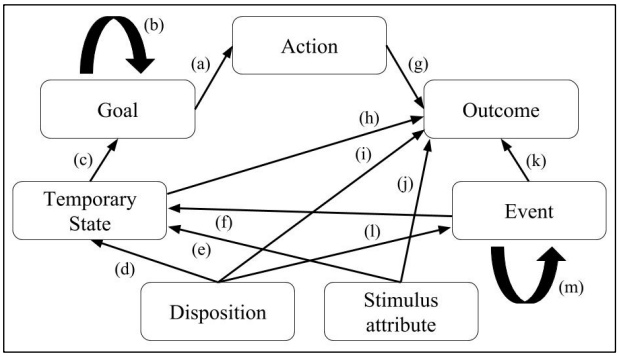


**Fig. 1.** Behavior and explanation types (boxes) and inference rules (arrows) in the Causal Explanation Network (CEN) model. Adapted from Böhm & Pfister [5].

which are assumed to reflect relations between behaviors and causal explanations (see arrows in Fig. 1).

## 2 Method

Sixty university students (mean age = 23, *SD* = 3.5 yrs., 80% women) individually completed a survey concerning how people interpret and explain behaviors. Participants in one experimental condition were presented with images of human behavior while participants in a second condition were presented with images of humanoid robot behavior (cf. Fig 2). The selection of behaviors used as experimental stimuli was based on the CEN model which distinguishes between four behavior types: *actions*, *outcomes*, *states*, and *events* [5]. To deal with positivity bias, i.e., the tendency to make different attributions for positive and negative events [6], one desirable and one undesirable behavior were presented for each behavior type. This resulted in eight individual behavior stimuli: positive action (A+): "Ellis mops the floor", negative action (A−): "Ellis lies about having cooked the dinner", positive outcome (O+): "Ellis makes a fantastic cake", negative outcome (O−): "Ellis burns the cake", positive event (E+): "Ellis gets tipped by the dinner guests", negative event (E−): "Ellis breaks a glass", positive state (S+): "Ellis is happy to be in the kitchen", negative state (S−): "Ellis is frustrated over cooking".

**Fig. 2.** Stimuli for the negative outcome, "Ellis burns the cake", in human and robot conditions.

Participants were asked to judge the plausibility of seven causes as explanations for the eight behaviors on a 7-degree ordinal scale ranging from "not at all" to "completely". The seven causal explanation types were derived from the CEN model and were given in the form *"Rate how plausible it is that the cause of Ellis' behavior is X"*, where X could be a *conscious goal,* an *action,* an *outcome*, an *uncontrollable event,* a *temporary state* (psychological or physical), a *disposition, or* an *attribute of someone or something in Ellis' environment*. The experimental conditions did not statistically significantly differ with regard to participant age, gender or self-assessed technical competence. Each participant gave written informed consent prior to participation.

## 3 Results

Participants' plausibility-ratings of behavior explanations statistically significantly differed between conditions in 8 out of 58 cases, as shown in **Table 1**. We note that *disposition* was rated as a more plausible cause for human behavior than humanoid behavior in 5 out of 8 cases, and that there were no statistically significant differences between participants' ratings of the positive event and negative action when enacted by the human versus the robot. Kendall's $W$ was run to assess inter-rater reliability of participants' judgements in human and robot conditions. Agreement was higher in the human condition, $W = .391, p < .0005$, than in the robot condition, $W = .294, p < .0005$.

**Table 1.** Judged plausibility of causal explanation types (rows) for human (left value in cell) and robot (right value in cell) behaviors (columns) with statistically significant differences at $p < .05$ marked in inverted colors.

| | **A+** | **O+** | **E+** | **S+** | **A−** | **O−** | **E−** | **S−** |
|---|---|---|---|---|---|---|---|---|
| **Goal** | 5.9/6.0 | 6.6/5.8 | 5.3/5.4 | 5.1/4.8 | 5.9/5.4 | 1.5/2.3 | 1.5/2.4 | 2.4/3.3 |
| **Action** | 6.2/5.9 | 6.0/5.7 | 4.7/5.3 | 5.0/4.6 | 5.6/5.5 | 4.7/4.3 | 4.7/4.9 | 4.8/4.0 |
| **Outcome** | 5.5/5.3 | 6.1/5.7 | 6.0/5.8 | 5.3/5.0 | 5.8/5.5 | 5.9/5.6 | 5.8/5.3 | 5.9/5.7 |
| **Event** | 2.7/2.3 | 2.3/2.1 | 3.4/3.0 | 2.6/3.3 | 2.3/2.9 | 4.1/3.7 | 4.7/3.7 | 4.6/3.7 |
| **Temp. state** | 2.9/3.0 | 2.7/3.3 | 3.1/3.1 | 4.8/4.3 | 3.9/3.9 | 3.1/3.8 | 3.8/3.5 | 5.1/4.6 |
| **Disposition** | 3.6/3.3 | 4.5/3.1 | 4.2/4.0 | 5.2/4.0 | 5.1/4.8 | 3.5/2.4 | 4.0/2.7 | 4.2/3.3 |
| **Stimulus attr.** | 3.4/4.9 | 4.0/4.0 | 5.2/5.1 | 3.8/4.6 | 3.7/4.5 | 3.3/4.0 | 3.5/3.7 | 3.7/4.5 |

## 4 Conclusion & Discussion

The seven causal explanation types were judged to be similarly plausible for human and robot behaviors in 50 out of 58 cases. This indicates that people perceive the causes of human and humanoid robot behavior similarly, at least given the particular participant demographic, scenario, human actor and robot featured in the study presented here. This can be taken to suggest that people rely on a common-sense concep-

tual framework of mind in judgments of robot behavior similar to those that has been studied and modeled in the human case [e.g., 4, 5, 7].

There were, however, a few notable differences between the two conditions. Firstly, we observed a trend in the data indicating that participants judged *disposition*s as generally more likely to be causes of human behavior. Robots were, for example, considered less likely to have dispositions that would cause them to be happy to be in the kitchen, to make fantastic cake, to burn a cake, or to break glasses. This raises the question whether people think of robots as less likely to have dispositions in the human sense, or as having less stable dispositions as humans, or whether people see robot dispositions as less efficacious in causing behavior than human dispositions. Secondly, we observed lower agreement among the participants judging causes of robot behaviors compared to judgements of human behavior. This might be taken to suggest that the influence of a shared folk-psychological framework is weaker in people's judgements of artificial behavior as compared to human behavior. While the present study does not provide answers as to why this effect occurred we suggest this as a future research topic.

We hope that the relatively simple methodology used in this initial study of human versus humanoid behavior can be applied and developed further to contribute to addressing the broader and more fundamental question of how people's social interactions with different types of both natural and artificial autonomous agents – such as robots, virtual agents, or automated cars – are effected by folk-psychological causal explanations of observed behavior, and to what degree the underlying mechanisms overlap or differ for different types of natural and artificial agents.

## References


1. Sciutti, A., Ansuini, C., Becchio, C., & Sandini, G.: Investigating the ability to read others' intentions using humanoid robots. *Frontiers in psychology*, *6*, 1362 (2015)
2. Wykowska, A., Chaminade, T., & Cheng, G.: Embodied artificial agents for understanding human social cognition. *Phil. Trans. R. Soc. B*, *371*(1693), 20150375 (2016)
3. Dennett, D. C.: Intentional systems. *The Journal of Philosophy, 68*(4), 87-106 (1971)
4. Heider, F.: The Psychology of Interpersonal Relations. Mansfield Center, CT, Martino Publishing (1958)
5. Böhm, G., & Pfister, H. R.: How people explain their own and others' behavior: a theory of lay causal explanations. *Frontiers in psychology, 6*, 139 (2015)
6. Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L.: Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological bulletin, 130*(5), 711 (2004)
7. Malle, B. F., & Knobe, J.: The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*(2), 101-121 (1997)