Differences in the Intentionality Bias when Judging Human and Robotic Action

Extended Abstract[†]

Antonia Eisenkoeck
Department of Psychology
Goldsmiths, University of London
SE14 6NW London
UK
aeise010@gold.ac.uk

James Moore
Department of Psychology
Goldsmiths, University of London
SE14 6NW London
UK
j.moore@gold.ac.uk

ABSTRACT

As social agents humans constantly make judgments about the behaviour of others. When observing ambiguous human actions humans are biased towards judging them as intentional, which is known as the intentionality bias. We compared intentionality bias scores when judging human and robotic behaviour to see if the intentionality bias was specific to human behaviour. Our results showed that for ambiguous but prototypically accidental scenarios participants in the robot condition were more likely to attribute intent than participants in the human condition. Groups did not differ significantly in age, gender, empathy or anthropomorphism. Additionally, anthropomorphism scores did not correlate with intentionality bias scores. Future work will explore if the reason for the observed higher intentionality bias scores is our expectation of robots being "error-free" and if varying the agent's humanness has an effect on judgements.

KEYWORDS

Intentionality bias, Intention attribution, Intention recognition, Human-Robot interaction, Social cognition

1 MAIN RESEARCH QUESTION

Intention attribution is a key element of social interaction. We constantly draw conclusions from other people's behaviour. Although it may feel like we judge objectively and only make informed decisions, past research suggests we are biased towards judging ambiguous human behaviour as intentional, known as the intentionality bias [2, 4]. As robotic agents are becoming increasingly integrated into our everyday life, we aim to find out if this bias is also present when judging robotic behaviour. We aim to explore whether strategies we apply or cognitive biases we show persist when interpreting robotic behaviour. Do more controlled cognitive processes, involving reasoning and inhibiting automatic judgements, make us reject the idea that a robot could act intentionally or are we similarly biased towards perceiving ambiguous behaviour as intentional?

2 BACKGROUND AND RELATED WORK

Understanding and attributing intent to observed behaviour is a key element of social cognition. It shapes how we interact with each other, as we react differently depending on whether we consider an action intentional or unintentional. For example, one might accept an apology from somebody who stepped on one's foot by accident but avoid somebody who did so intentionally. Unlike this example, our social world is rather complex and intentionality is not always clear. Therefore, in so-called ambiguous situations, we sometimes have to guess whether an action was intentional or not. One might assume that in truly ambiguous situations, in which an action could have been done on purpose or by accident, we would be as likely to judge it intentional as we would unintentional. However, as the intentionality bias suggests, we tend to automatically attribute intent to ambiguous behaviour [4]. This bias has been consistently found in studies of human action, using verbal [4] as well as visual [2] paradigms. One possible explanation is that judging behaviour to be intentional is cognitively less demanding, because it does not require an alternative cause [1]. Humans are interacting with artificial intelligence more regularly and robots are becoming incorporated into our social thinking. Therefore, it is important to explore whether we apply similar cognitive processes when judging robotic behaviour as when judging human behaviour. If we do, and our automatic tendency to attribute intent withstands more controlled processes suggesting that a robot cannot act intentionally, one would expect similar patterns when judging robotic and human behaviour. If, however, cognitive processes involved in intention attribution differ when observing human and robotic actions, one would expect intentionality judgments to differ accordingly. Intuitively, one might assume that intentional actions require a mental state or mental activity, as it is inherent to a plan, and that artificial intelligence does not give rise to what is commonly understood as a mental state. In this case, we would expect people to hardly ever attribute intent to a robotic action.

We aim to explore this question in a recent study outlined below.

3 RESEARCH METHODOLOGY

3.1 Design and Participants

In an online study we asked 133 participants to complete a modified version of Rosset's (2008) intentionality bias paradigm as well as two questionnaires (measures described below) [4]. Participants were healthy volunteers between 18 and 50 years old. A between-groups design was applied.

3.2 Intentionality Bias Paradigm

To measure the intentionality bias, we used a modified version of Rosset's sentence paradigm [4]. The initial paradigm consists of 34 test sentences describing ambiguous human actions, 12 of which are prototypically accidental (example: "He broke the window."), and 22 are prototypically intentional (example: "She made a mark on the paper."). Participants are asked to judge if the action was done "on purpose" or "by accident". In our modified version of the paradigm each scenario described an action performed by "R." (example: "R. broke the vase."). For one group, "R." referred to Robert, a human, and for the other group "R." referred to a robot. A short vignette at the start of the experiment introduced "R." as Robert, a human, or as a robot (see A.1 for vignettes). Five sentences from the initial paradigm had to be excluded as they describe actions that a robot could not have carried out (see A.2 for a full list of the sentences used). Participants' intentionality bias scores (IS) were calculated as the percentage of "on purpose" judgments of items responded to for both types of sentences separately, prototypically intentional (PI) and prototypically accidental (PA), resulting in a PI-IS and a PA-IS for each participant.

3.3 Anthropomorphism Measure (IDAQ)

In addition to the PI-IS and PA-IS, we also measured participants' tendency to anthropomorphise using the Individual Differences in Anthropomorphism Questionnaire (IDAQ) [6], as it was hypothesised that the tendency to anthropomorphise might be related to viewing the robot as an intentional agent.

3.4 Empathy Measure (QCAE)

As empathy might play a role in how we understand and perceive intentional behaviour, the Questionnaire of Cognitive and Affective Empathy (QCAE) was also administered, which allowed us to see if groups differed in cognitive or affective empathy [3].

4 RESULTS

For the analysis we only included participants who had responded to at least 75% of the items of the corresponding measure. One participant had to be excluded because they were diagnosed with autism spectrum disorder, which is associated with atypical intention attribution patterns.

4.1 Prototypically Accidental Scenarios

An independent-samples t-test was conducted on data from 111 participants and showed that there was a significant difference between PA-ISs in the human and robot condition (t(109) = -3.61, p<.001). For prototypically accidental scenarios, participants in the robot condition were more likely to attribute intent to actions than participants in the human condition (Table 1).

4.2 Prototypically Intentional Scenarios

An independent-samples t-test was conducted on data from 98 participants and revealed no significant difference between PI-ISs in the human and robot condition (t(96)=0.417, p=0.678). For prototypically intentional scenarios, participants of both groups were as likely to attribute intent to actions (Table 1).

Table 1: Mean and PA-IS and PI-IS and standard deviation for human and robot condition

	Human	Robot
PA-IS	M=19.27 *(SD=13.88)*	M=30.21 *(SD=18.05)*
PI-IS	M=71.02 *(SD=23)*	M=69.18 *(SD=18.95)*

4.3 IDAQ, QCAE and Demographics

There were no significant group differences in age, gender, IDAQ or QCAE scores. Hence, we judge the groups to be homogenous in these aspects. Furthermore, IDAQ scores did not significantly correlate with PA-ISs and PI-ISs respectively in either group (p>.05). This suggests anthropomorphic tendencies did not influence intentionality judgements in either type of ambiguous actions.

5 DISCUSSION AND FUTURE WORK

The results suggest that people are as likely to attribute intent to ambiguous but prototypically intentional actions when the action is performed by a human as when it is performed by a robot. Both groups judged behaviour as intentional in most scenarios.

However, in situations involving ambiguous but prototypically accidental action people are more likely to attribute intent when the agent is a robot than when it is a human.

The main outstanding question is why participants in the robot condition were more likely to attribute intent to ambiguous but PA scenarios. All PA sentences describe an action with a negative outcome, i.e. if done by accident the action can be understood as a mistake, which the agent would usually have avoided if aware of what they were doing. Previous research suggests that we perceive making mistakes as a uniquely human quality [5]. Artificial intelligence, on the other hand, is possibly expected not to make any mistakes but be "error-free". This potentially explains why participants in the robot condition were less likely to judge the robot's behaviour as unintentional: they did not expect it to commit a mistake but act according to its intentions or goals.

In a follow-up study, we plan to manipulate the error-proneness of the robot by changing the vignettes so that the robot never, sometimes or often makes mistakes. If people attribute intent to robotic action because they do not expect robots to make mistakes, manipulating error-proneness should have an effect on intentionality bias scores in prototypically accidental situations.

Additionally, to examine the effect of humanness on observers' intentionality bias, we will vary the agent's degree of humanness in future studies. This will be achieved by supplementing the vignette introducing the agent with a picture of a human face, a humanoid robot and a non-humanoid robot. Our prediction is that intentionality bias scores will be lower the more human the physical appearance of the agent.

In summary, our results suggest differences in intentionality judgments for human and robotic behaviour. We seem to be less likely to perceive robotic behaviour as accidental or erroneous. A better understanding of these differences will help to adjust their design. As increasingly common social agents, robots' characteristics will influence how we interact with them and, hence, have an impact on our social lives.

APPENDICES

A.1 Vignettes

A.1.1 Vignette Human Condition In this section of the study, you will read a series of sentences describing an action by Robert (from now on referred to as R.). R. operates as an aid around the house. During the course of a day R. does various jobs and a lot happens. Your task is to decide whether the actions described in the following sentences were done "on purpose" or "by accident" by checking the appropriate box.

A.1.2 Vignette Robot Condition. In this section of the study, you will read a series of sentences describing an action by a robot (from now on referred to as R.). R. operates as an aid around the house. During the course of a day R. does various jobs and a lot happens. Your task is to decide whether the actions described in the following sentences were done "on purpose" or "by accident" by checking the appropriate box.

A.2 Vignettes

A.2.1 Prototypically Accidental Sentences

- R. hit the man with his car.
- R. burnt the meal.
- R. broke the vase.
- R. tracked mud inside.
- R. forgot the homework.
- R. arrived 5 minutes late.
- R. broke the window.
- R. woke the baby up.
- R. stepped in the puddle.
- R. set off the alarm.
- R. dripped paint on the canvas.
- R. kicked the dog.
- R. left the water running.
- R. set the house on fire.
- R. popped the balloon.

A.2.2 Prototypically Intentional Sentences

- R. ripped the piece of paper.
- R. cut him off driving.
- R. deleted the email.
- R. drove over the speed limit.
- R. ignored the question.
- R. knocked over the sand castle
- R. made a mark on the paper.
- R. sprayed him with water.
- R. took an illegal left turn.
- R. walked by without saying hello.
- R. left without leaving a tip.

ACKNOWLEDGMENTS

This work was supported by a Leverhulme Trust Research Project Grant to James Moore (RPG-2016-012).

REFERENCES

[1] L. Bègue, B.J. Bushman, P.R. Giancola, B. Subra and E. Rosset, 2010. "There is no such thing as an accident," especially when people are drunk. *Personality and social psychology bulletin, 36*(10), pp.1301-1304.

[2] J.W. Moore and A. Pope, 2014. The intentionality bias and schizotypy. The Quarterly *Journal of Experimental Psychology, 67*(11), pp.2218-2224.

[3] R.L. Reniers, R. Corcoran, R. Drake, N.M. Shryane, and B.A.Völlm, 2011. The QCAE: A questionnaire of cognitive and affective empathy. *Journal of personality assessment*, 93(1), pp.84-95.

[4] E. Rosset, 2008. It's no accident: Our bias for intentional explanations. *Cognition, 108*(3), pp.771-780.

[5] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, F. and Joublin, 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics, 5*(3), pp.313-323.

[6] A. Waytz, J. Cacioppo, and N. Epley, 2010. Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), pp.219-232.