Studying the Craft of Folk Psychology in HRI

Sam Thellman & Tom Ziemke
Dept. of Computer and Information Science, Linköping University
sam.thellman@liu.se, tom.ziemke@liu.se

Abstract

Human interaction with intelligent autonomous systems depends on the interpretation of behavior in terms of mental states. However, studies of mental state attribution to robots have so far focused primarily on folk theories about robots ("How do people think about the mental states of robots?") without considering the function of mental state attribution in human–robot interactions. This paper highlights a number of limitations to this approach and argues the importance of studying: (1) robots in ecologically valid contexts, (2) how specific attributions affect people's ability to predict and explain robot behavior, and (3) what causes people to attribute specific kinds of mental states to robots. Two novel methodological approaches are proposed and discussed.

1 Introduction

A central – but commonly overlooked – question in the design of human multimodal interaction with intelligent autonomous systems, such as social robots or automated vehicles, is how, when and why people interpret the behavior of autonomous systems in terms of (attributed) mental states, and how this affects their ability to interact with such systems. People's ability to predict and explain behavior (human and non-human alike) depends crucially on such attribution processes [e.g., Dennett, 1989]. Partly for this reason, there has been a recent resurgence in the area of explainable artificial intelligence [Miller, 2017] and a growing interest in the role of intentions and intentionality amongst human–robot interaction researchers [Thill and Ziemke, 2017].

Previous studies of mental state attribution to robots have focused primarily on folk theories of robots ("How do people think about the mental states of robots?"). For example, Gray *et al.* [2007] found that people would attribute mental states related to agency (e.g., memory, planning, and thought) but not subjective experience (e.g., fear, pain, pleasure). Systma and Machery [2010] showed that people with and without philosophical training (as determined by responses to biographical survey questions) differ in their attributions, with non-philosophers being more inclined to attribute mental states to robots overall. Buckwalter and Phelan [2013, p. 353] argued that attributions are influenced by tacit assumptions about the intended function of robots and found in their experiments that "ordinary people (but not trained philosophers) are making different assumptions about the robot across different probes". This sentiment was partly echoed in a study by Fiala, Arico, and Nichols [2014, p. 37] which investigated whether the design used in earlier experiments of using forced-choice questions in conjunction with images and/or descriptions of robots "puts undue pressure on subjects to attribute certain states to robots". In their experiments it was found that respondents – when allowed to choose between different ways of describing the capabilities of a robot (e.g., the robot "detected green" vs. "saw green" or "identified the location of the box" vs. "knew the location of the box") – preferred not to attribute mental states at all. The authors concluded that:

> When we probe people for their explicit judgments about whether robots have mental states, responses are influenced by a wide variety of factors. The apparent function of the robot, the nature of the question (forced choice vs. not), and platitudes about robots may all contribute to producing reasoned judgment about the state of robots. [Fiala *et al.*, 2014, p. 44]

There is so far little agreement about what kinds of mental states people attribute to robots, and the commonly adopted methodology of asking people to directly judge mental states of robots is subject to a number of limitations. In addition to the limitations highlighted above, research by Thellman and Ziemke [2017] indicates that people's responses to questions regarding their views on robots are easily manipulated. These considerations provoke the question "What can actually be learned from asking people directly about their thoughts about the mental states of robots?".

Fiala *et al.* [2014] suggest, as a way to overcome methodological limitations in previous research, to distinguish between deliberative "high-road" and automatic "low-road" attribution processes, where the latter refers to "a more fundamental tendency to treat robots as fully minded ... that tends to manifest in the form of automatic, unreasoned attribution of a wide range of mental states to robots" [Fiala *et al.*, 2017, p. 44]. The authors review several examples from the empirical literature that illustrate how low-road processes come into

10.21437/AI-MHRI.2018-7

play in people's attributions to robots, and argue that people often refrain from attributing mental states to robots because of a high-road process that invokes "the culturally prevalent platitude that robots do not have minds" [Fiala *et al.*, 2017, p. 44]. However, the authors provide no suggestion on how to study low-road attribution processes experimentally.

In contrast to this approach, the present paper proposes to overcome limitations of previous studies by sticking with the high-road (probing people's explicit attributions to robots) while introducing increased experimental control. Two alternative methodological approaches are suggested. The first approach, taken in our empirical research so far [Thellman *et al.*, 2017b; Petrovych *et al.*, in press; Thellman *et al.*, 2017a], involves studying mental state attribution to humans versus robots in ecologically valid contexts where the function of the robot is more clearly conveyed (section 3). The second approach is to study what philosopher Daniel Dennett [1991] refers to as the "craft" of people's folk-psychological theories about robots by controlling experimentally for the causes and effects of mental state attributions to robots in the context of human–robot interactions (section 4). The paper concludes with a brief summary of the position taken by the authors on the way forward for studying mental state attribution to robots in HRI (section 5).

2 The Craft of Folk Psychology

Daniel Dennett [1991] calls what people actually do in folk psychology (i.e., predicting and explaining behavior based on attributed mental states) *the craft* and distinguishes this from how people talk about what they do, which he calls *the theory*. Folk theory is occasionally referred to as "the ideology" to emphasize that people's folk-psychological notions about persons and other things are not necessarily true. This includes the attribution of specific beliefs, desires, emotions, qualitative and phenomenological states, and the idea that these things are (at least in the human case) located *in the head somewhere*. As suggested by Fiala *et al.* [2014], cultural platitudes such as that *robots do not have minds* might also be part of people's folk theories about robots.

The craft of folk psychology can be understood as the practice of using folk psychology as a predictive strategy by attributing behavior to underlying mental states and processes. Dennett refers to this practice as adopting *the intentional stance* [Dennett, 1989]. When attributing mental states (and rationality) to an entity, one draws upon information relevant to the possession of specific goals, constraints and beliefs, such as social roles, morphology, environment, sensory capabilities, etc. We will refer to such information as *attribution causes*. The attribution of a specific goal, constraint, or desire (or a collection of these things) gives rise to specific behavior predictions and explanations. We will refer to predictions and explanations as *attribution effects*. On this view, the "craft" of folk psychology is understood as comprising three components: causes, attribution (i.e., an element of folk theory), and effects (see Figure 1).

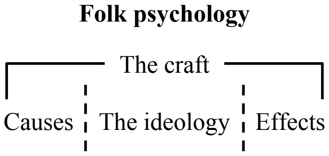


Figure 1: Previous studies of mental state attribution to robots have focused primarily on folk theories of robots ("the ideology") independently of how they are used ("the craft").

One of Dennett's key insights about the craft of folk-psychology is that people's ability to predict and explain behavior is logically independent from their talk about behavior:

> Whether one calls what one ascribes to the [entity] beliefs or belief-analogues or information complexes or Intentional whatnots makes no difference to the nature of the calculation one makes on the basis of the ascription. [Dennett, 1971, p. 91]

The ideology does not have to be "true" in order for people's predictions and explanations to be *consistent* with the behavior observed. The behavior of an entity can in many cases be consistently explained and predicted based on different kinds of ascriptions, and different ascriptions can be arrived at based on the same considerations (see Figure 2). Hence, research on mental state attributions to robots will continue to be uninformative about the role of mental state attribution in shaping people's predictions and explanations of the behavior of specific robots, as long as specific attributions are not analyzed in terms of function (i.e., causes and effects).

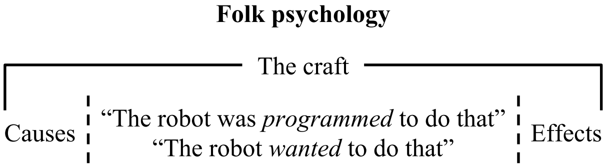
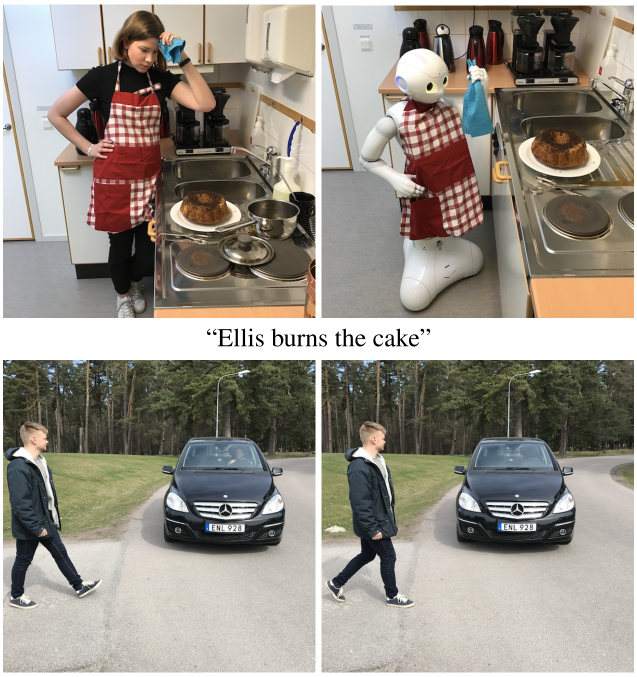

Folk psychology
The craft
Causes
"The robot was *programmed* to do that"
"The robot *wanted* to do that"
Effects

Figure 2: Intentional and non-intentional attributions can have identical causes and effects, i.e., they can be functionally equivalent.

3 Human versus Robot Behavior Comparisons

Thellman *et al.* [2017a] exposed experiment participants to images and verbal descriptions of different behaviors exhibited either by a person or a humanoid robot in the context of a kitchen scenario (Figure 1, top). Participants were asked to rate the intentionality, controllability and desirability of the behaviors and to judge the plausibility of seven different types of explanations. The results indicated that attributions to the human and the robot were highly similar, with some minor differences. Adopting the same methodology to investigate people's interpretations of human-driven vs. driverless car behaviors in road traffic scenarios (Figure 1, bottom), Petrovych *et al.* [in press] had similar results but also found considerably lower agreement in participant ratings of the driverless behaviors.

As pointed out by Buckwalter and Phelan [2013], attributions to robots seem to vary depending on people's tacit as-



“Ellis burns the cake”

“The car stops to late at the unmarked crossing”

Figure 3: Examples of stimuli used in comparisons of mental state attribution to a human vs. humanoid robot [Thellman *et al.*, 2017] and a human-driven vs. self-driving car [Petrovych *et al.*, forthcoming] in depicted real-life interaction scenarios.

sumptions about the function of the robot in question. This approach mitigates this concern as stimuli are based on real-life scenarios that convey the intended function of the robot relatively clearly (assuming that people expect the function of the robot to be similar or identical to the function of a person in the same scenario). Another strength of the approach is that it may raise the ecological validity of the study results. Human–robot comparisons are interesting in their own right because robots are in many cases expected to replace human labor such as driving, with equal or better performance than humans. Studying people’s attributions to robots featured in real-life scenarios may therefore yield results that are more relevant to real-life applications than using hypothetical examples of robots where the function of the robot is unclear or unspecified [cf. Gray *et al.*, 2007; Sytsma and Machery, 2010; Buckwalter and Phelan, 2013; Fiala *et al.*, 2014].

A significant limitation of the methodology used in the comparative studies mentioned above, which concerns all studies on mental state attribution to robots conducted so far, is that it is unclear what to make of the results. What does it mean that a person ascribes a specific mental state to a robot? And how can such results be used to improve human–robot interactions? In the following section it is argued that research on mental state attribution will remain largely uninformative with regard to these questions unless it starts to consider how people use mental state attribution in human–robot interactions (hence, “studying the craft of folk psychology in HRI”).

4 Measuring Causes and Effects of Mental State Attribution to Robots

In their pioneering experiments on attribution, Heider and Simmel [1944] asked participants to interpret an animated short-film featuring three geometrical figures (Figure 4). In one condition, participants were given the general instruction to *write down what happened in the picture* and in another condition participants were asked more specific questions such as *In one part of the movie the big triangle and the circle was in the house together. What did the big triangle do then? Why?* The main finding was that all but a single participant interpreted the film in terms of mental states, such as beliefs, desires and intentions (regardless of instructions and questions asked). Heider and Simmel [1944, p. 259] concluded that “this method is useful in investigating the way the behavior of other persons is perceived”.

Three advantages of the methodology used in Heider and Simmel’s experiments stand out. Firstly, the use of video material allows the experimenter to convey the dynamics of interpersonal interactions to experiment participants while retaining a level of experimental control which can be difficult or impossible to achieve using real-time enactions of behavior (including robot behavior) in complex real-world scenarios. Secondly, the open-ended approach of asking people to freely explain behavior mitigates the concerns raised by Fiala *et al.* [2014] over putting undue pressure on participants to attribute specific mental states (e.g., forced-choice questions). Finally, the third and perhaps most significant methodological advantage – one which Heider and Simmel do not take advantage of in their experiments – is that their approach allows for measuring the causes (“What causes people to attribute specific kinds of mental states?”) and effects (“How do specific attributions affect people’s ability to predict and explain robot behavior?”) of mental state attributions in (e.g., human–robot) interactions.

Most importantly, for the purposes of studying mental state

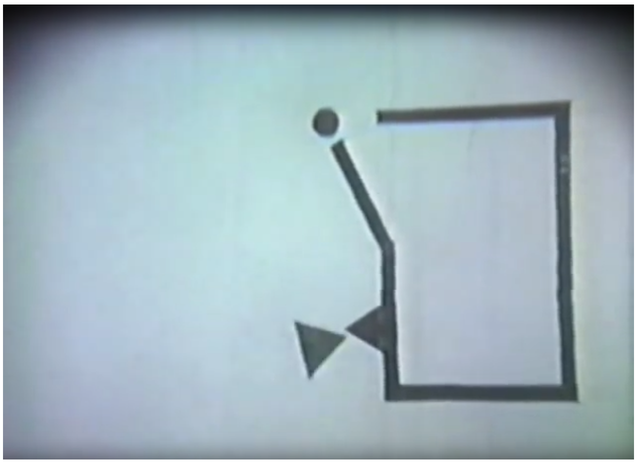


Figure 4: “The small circle gets out of the way as the triangles fight”, according to one of the participants that interpreted a scene from an animated short-film in the context of Heider & Simmel’s [1944] experiments.

attribution to robots, a video sequence featuring human–robot interactions can be paused to allow the participants to make predictions of *What will happen next?* or *What will the robot do next?* Prediction outcomes can be measured (i.e., one can measure, for instance, whether the participant succeeded or failed to predict the self-driving car crash). The socratic-method-style why-questions employed by Heider and Simmel can be used to probe people's motivations for their (failed or successful) predictions (e.g., the participant failed to predict the self-driving car crash because she thought that the car could see the other driver). These answers can be followed-up (perhaps best using interview methodology) with questions regarding the basis of the misinterpretations (e.g., the participant thought that the car could see the other driver because she thought that the range of the car's "sight" was coextensive with the area illuminated by the car's headlights).

To sum up, using video stimuli featuring human-robot interactions, the experimenter can ask participants by way of questionnaire or interview methodology to predict and explain behavior and to motivate their predictions and explanations. This method allows HRI researchers to study the craft of folk psychology in the context of human–robot interactions, that is, what people *do* with mental state attributions to robots (or what mental state attributions do for them), in contrast to studying attributions in isolation from their use.

5 Conclusion

This paper has highlighted a number of limitations in previous studies of mental state attribution to robots. Firstly, most previous studies are based on hypothetical scenarios that leave the function of the robot unclear or unspecified. A comparative approach, featuring human versus robot behaviors in real-life scenarios as stimuli, was suggested to mitigate this issue. Secondly, previous studies have so far ignored the function of mental state attributions to robots. In short, they have not considered the "craft" of people's folk psychological theories about robots [Dennett, 1991]. This includes investigating the effects that the attribution of specific mental states have on people's ability to predict and explain the behavior of robots, and investigating the causes of attributions. It was argued based on Dennett's theories that very little of practical value can be said about human–robot interactions based on studies of mental state attributions *unless* the function (craft) of specific attributions is considered. In particular, as long as the typical and idiosyncratic causes of misinterpretation of robot behavior remains unknown it is difficult or impossible to design robots so as to reduce misguided explanation and prediction of robot behavior. Finally, based on the pioneering work of Heider and Simmel [1944] which employed short videos of interpersonal interactions, an experimental approach was proposed as a way forward to study the function of mental state attribution in human interactions with autonomous systems.

References

[Buckwalter and Phelan, 2013] Wesley Buckwalter and Mark Phelan. Function and feeling machines: a defense of the philosophical conception of subjective experience. *Philosophical Studies*, 166(2):349–361, 2013.

[Dennett, 1971] Daniel Dennett. Intentional systems. *The Journal of Philosophy*, 68(4):87–106, 1971.

[Dennett, 1989] Daniel Dennett. *The intentional stance*. MIT press, 1989.

[Dennett, 1991] Daniel Dennett. Two contrasts: folk craft versus folk science, and belief versus opinion. *The future of folk psychology: Intentionality and cognitive science*, pages 135–148, 1991.

[Fiala *et al.*, 2014] Brian Fiala, Adam Arico, and Shaun Nichols. You, robot. In Edouard Machery and Elizabeth O'Neill, editors, *Current controversies in experimental philosophy*, pages 31–47. Routledge, 2014.

[Gray *et al.*, 2007] Heather M Gray, Kurt Gray, and Daniel M Wegner. Dimensions of mind perception. *Science*, 315(5812):619–619, 2007.

[Heider and Simmel, 1944] Fritz Heider and Marianne Simmel. An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259, 1944.

[Miller, 2017] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269*, 2017.

[Petrovych *et al.*, in press] Veronika Petrovych, Sam Thellman, and Tom Ziemke. Interpretation of goal-directed autonomous car behavior. In *CogSci 2018: Changing Minds. 40th Annual Meeting of the Cognitive Science Society, Madison, VA*. Cognitive Science Society, in press.

[Sytsma and Machery, 2010] Justin Sytsma and Edouard Machery. Two conceptions of subjective experience. *Philosophical Studies*, 151(2):299–327, 2010.

[Thellman and Ziemke, 2017] Sam Thellman and Tom Ziemke. Social attitudes toward robots are easily manipulated. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 299–300. ACM, 2017.

[Thellman *et al.*, 2017a] Sam Thellman, Annika Silvervarg, and Tom Ziemke. Folk-psychological interpretation of human vs. humanoid robot behavior: exploring the intentional stance toward robots. *Frontiers in Psychology*, 8:1962, 2017.

[Thellman *et al.*, 2017b] Sam Thellman, Annika Silvervarg, and Tom Ziemke. Lay causal explanations of human vs. humanoid behavior. In *International Conference on Intelligent Virtual Agents*, pages 433–436. Springer, 2017.

[Thill and Ziemke, 2017] Serge Thill and Tom Ziemke. The role of intentions in human-robot interaction. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 427–428. ACM, 2017.