Artificial Intelligence for Human-Robot Interaction
AAAI Technical Report FS-17-01

How People Explain Action (and Autonomous Intelligent Systems Should Too)

Maartje M. A. de Graaf, Bertram F. Malle
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University
maartje_de_graaf@brown.edu & bfmalle@brown.edu

Abstract

To make Autonomous Intelligent Systems (AIS), such as virtual agents and embodied robots, "explainable" we need to understand how people respond to such systems and what expectations they have of them. Our thesis is that people will regard most AIS as intentional agents and apply the conceptual framework and psychological mechanisms of human behavior explanation to them. We present a well-supported theory of how people explain human behavior and sketch what it would take to implement the underlying framework of explanation in AIS. The benefits will be considerable: When an AIS is able to explain its behavior in ways that people find comprehensible, people are more likely to form correct mental models of such a system and calibrate their trust in the system.

Introduction

The call for Autonomous Intelligent Systems (AIS) to be transparent has recently become loud and clear (Wachter, Mittelstadt, and Floridi 2017). Some forms of transparency, such as traceability and verification, are particularly important for software and hardware engineers (Cleland-Huang, Gotel, and Zisman 2012; Fisher, Dennis, and Webster 2013); other forms, such as explainability or intelligibility, are particularly important for ordinary people. Explanation is arguably a three-value predicate: someone, a communicator, explains something to someone, an audience (Bromberger 1965; Hilton 1990). The success of an explanation therefore depends on several critical *audience factors*—assumptions, knowledge, and interests that an audience has when decoding the explanation. In this paper, we focus on ordinary people as the audience and on one fundamental audience factor: the conceptual and linguistic framework within which people explain human behavior. We propose that explainable AIS must generate explanations within the conceptual and linguistic bounds of this framework.

Autonomous Intelligent Systems

Autonomous Intelligent Systems (AIS) come in many forms: some are strikingly human-like robots that move in social spaces, others are virtual assistants communicating from a computer screen, and yet others are powerful algorithms implemented in deep neural networks. Robots and virtual agents typically have numerous features that trigger human inferences of *intentional agency* (e.g., eyes, self-propelled movement, contingent interaction) (Johnson 2000; Premack 1990). We propose that people's demands for explanation are most pressing for such salient intentional agents, especially when they interact and communicate with people in social contexts. Our analysis will therefore focus on AIS that people clearly treat as *agents* and that perform actions people consider *intentional* (e.g., making decisions, offering suggestions). For those intentional agents, we hypothesize, people will apply the same conceptual framework of behavior explanation that they apply to humans; and they will expect AIS to apply this framework as well. There may be a subset of AIS that people do not regard as intentional agents; and for those, they may apply a purely mechanical explanatory framework. We believe, however, that systems that are in fact autonomous and intelligent will almost always exhibit some indicators of intentional agency (e.g., initiative, planning, decision making), and as soon as these indicators lead people to actually regard them as intentional agents, people will apply the human conceptual framework of behavior explanation to them.

Human Conceptual Framework of Behavior Explanation

A core component of human social interaction involves the explanation of one's own and others' behaviors. Behavior explanations provide people with meaning and understand-

ing of the myriad of behaviors they encounter every day, and these explanations guide how people respond to, predict, and influence others' behaviors. Scholars from many disciplines have converged on the insight that people's ordinary behavior explanations are embedded in a fundamental conceptual framework, often called folk psychology or theory of mind (Heider 1958; Horgan and Woodward 1985; Premack and Woodruff 1978). Core concepts in this framework are *agent*, *intentionality*, and *mind*, and they are closely related (D'Andrade 1987; Malle 2005). Objects that appear self-propelled and behave contingently with the perceiver are taken to be agents (Johnson 2000; Premack 1990) . For such agents, the perceiver is sensitive to face, gaze, and motion cues that reveal which of the agent's behaviors are intentional (Dittrich and Lea 1994; Phillips, Wellman, and Spelke 2002). And these intentional behaviors perceived as caused by key mental states, such as beliefs, desires, and intentions (Malle and Knobe 1997a; Searle 1983).

Many studies have shown that people regard AIS as agents, treat many of their behaviors as intentional, and infer mental states from those behaviors (Harbers, van den Bosch, and Meyer 2009; Levin et al. 2013; Monroe, Dillon, and Malle 2014; Voiklis et al. 2016). It is only a small step to posit that people will explain behaviors of an AIS using the same conceptual framework they use to explain human behaviors. Moreover, people are likely to expect other people to also explain the AIS behavior in this way. It is a slightly larger but still reasonable step to posit that when people wonder why an AIS acted a certain way, they will expect the AIS to explain its own behavior using that same framework of agency, intentionality, and mind.

In this paper, we briefly introduce past work on how people generally make sense of AIS, especially robots, then turn to our theoretically and experimentally grounded analysis of how people generally explain human behavior, and then develop implications of this analysis for how to implement explanations in AIS such as robots. In this way, we provide a psychologically grounded frame for what it would take for AIS to be "explainable" to ordinary human beings.

Making Sense of Autonomous Intelligent Systems

A considerable amount of previous work has focused on making AIS, and robots specifically, expressive and socially aware (Hoffman et al. 2014; Huang and Mutlu 2012; Triebel et al. 2016). Although this has improved social engagement between humans and such systems, it does not necessarily improve the transparency of the sequences of actions such systems perform. Several studies in human-robot interaction show that people readily use their experience of what people generally know to determine what robots "know" (Lee et al. 2010; Powers et al. 2005). Indeed, when people interact with an AIS, they will inevitably construct mental models to understand and predict its actions and lower their uncertainty experienced during interactions (Epley, Waytz, and Cacioppo 2007). However, people's mental models of AIS stem from their interactions with living beings. Thus, people easily run the risk of establishing incorrect or inadequate mental models of artificial systems, which could result in self-deception or even harm (Wortham, Theodorou, and Bryson 2016a). Moreover, a long-term study (Tullio et al. 2007) showed that initially established (incorrect) mental models of an intelligent information system remained robust over time, even when details of the system's implementation were given and initial beliefs were challenged with contradictory evidence.

Incorrect mental models of AIS can have significant consequences for trust in such systems and, as a result, for acceptance of and collaboration with these systems (Wang, Pynadath, and Hill 2016). Several studies indicate that people distrust an AIS when they are unable to understand its actions. When a robot fails to communicate its intentions, people not only perceive that robot as creepy or unsettling (Williams, Briggs, and Scheutz 2015) they also perceive such robots as erratic and untrustworthy even when they follow a clear decision-making process (Lomas et al. 2012). Indeed, when a robot is not transparent about its intentions (i.e., not providing any explanations for its behavior), people may even question its correct task performance and blame the agent for its alleged errors (Kim and Hinds 2006). In addition to such cases of distrust, incorrect mental models of AIS can also lead to the opposite situation. People sometimes over-trust artificial agents, such as when they comply with a faulty robot's unusual requests (Salem et al. 2015) or follow the lead of a potentially dysfunctional robot (Robinette et al. 2016).

Under what conditions, then, will people have calibrated trust in AIS? One important condition is accurate knowledge of the AIS's abilities and appropriate domains of performance; another condition is accurate knowledge of the AIS's beliefs, goals, and plans (i.e., its "mental states"). When a robot provides explanations of its own actions, people gather more reliable information about abilities and mental states (Kiesler 2005; Mueller 2016) and are therefore able to build more accurate models of that robot (Wortham, Theodorou, and Bryson 2016b). As a result, people will correctly calibrate their trust in such systems (Theodorou, Wortham, and Bryson 2016; Wang, Pynadath, and Hill 2016)). Thus, there should be little doubt about the value of making robotic systems transparent and explainable, and indeed about the value of robots that explain their own decisions and behaviors.

Before addressing how to implement behavior explanations in AIS, we first offer an analysis of what such explanations should look like if they have the intended effect of increasing understanding, trust, and human-machine collaboration.

Ordinary Behavior Explanation

Many philosophers and psychologists agree that humans separate the entire realm of behavior into intentional and unintentional events (Heider 1958; Malle, Moses, and Baldwin 2001; Searle 1983). Likewise, many linguists count the concept of intentionality as fundamental to the way humans see the world, and linguistic forms of this concept have been found across all known languages (Bybee, Perkins, and Pagliuca 1994; Jackendoff and Culicover 2003). In short, the concept intentionality is the hub of people's conceptual framework of mind and behavior. We now sketch how this concept is constituted and how it lays the groundwork for people's explanations of behavior.

People show a high level of agreement in their judgments of events to be either "intentional" or "unintentional" (Malle and Knobe 1997a; Ohtsubo 2007). Intentional events are brought about by identifiable agents, and what makes the event intentional is that it is directly caused by the agent's reasoning and choice. Specifically, people's concept of intentional action is constituted as follows (Malle and Knobe 1997a, 2001): Agents intentionally perform action *A* when they have an intention to *A*, have the skill to *A*, and have awareness of *A*-ing while *A*-ing. In addition, they form the intention to *A* on the basis of a desire for an outcome *O* and the belief that *A* leads to *O*. All other events are "unintentional," lacking the critical role of an intention as mediating between the mind and the action (Heider 1958).

How Do People Explain Intentional Behaviors?

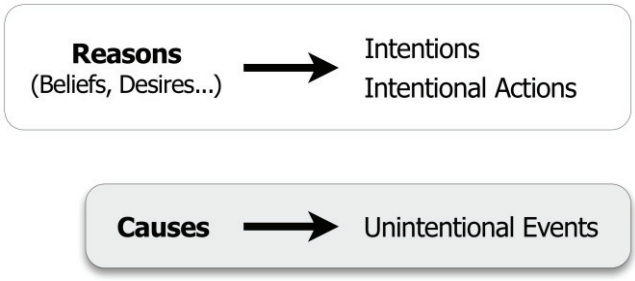
The concept of intentionality and its critical components of belief, desire, and intention lay the foundation for people's folk explanations of behavior, which come in distinct modes (Malle 1999, 2004, 2011). The primary mode of explaining intentional behaviors is by way of *reason explanations*. These cite an agent's reasons for intending to act or for acting intentionally. Reasons refer to the desire and belief components of intentionality but also often include informative background beliefs or desires in the context of which the central belief-desire-intention reasoning occurred. Here are two examples:

(1) "Why did you start jogging?" – "Because I wanted to get in better shape [desire], and … I figured that jogging is going to help [belief]."

(2) She told me to stay away from the neighbors' kids because she knew they were a bad influence on me [belief].

Unintentional behaviors, by contrast, do not stem from intentions and belief-desire reasoning; instead, they are explained by reference to a wide variety of causes, such as physiology ("She felt unwell because she had an infection"), behaviors ("He tripped on the last 100 meters because his opponent hit his knee"), or culture ("They were devastated by the slight because you don't treat people that way in their country"). Such *cause explanations* of unintentional behavior are conceptually no different from cause explanations of physical events such as rolling rocks or computers running out of battery; merely the kinds of causes differ. In none of these cases does anybody or anything form an intention or make a decision (people don't intend or decide to be devastated; computers don't intend or decide to run out of battery).

Sometimes people explain intentional actions using two other modes (causal history of reason explanations and enabling factor explanations), either in addition to or instead of reason explanations. For further discussion of these modes of explanation, see (Malle 2011; Malle et al. 2000; McClure and Hilton 1998); here we focus on the two



contrasting modes of reason explanations for intentions/intentional actions and cause explanations for unintentional events (see Figure 1).

Figure 1. Basic schematic of two distinct modes of explanation people apply to intentional and unintentional events

Reason Explanations for Intentional Behaviors

Naturally, then, when we examine people's explanations of AIS behaviors and when we design AIS to explain their own behaviors, reason explanations of intentional action are of prime interest. An AIS's technical features, malfunction, or unintended effects require reference to causes in software, electronics, or physics; intentional actions will—if people treat AIS largely like human agents—require reference to the unique category of "reasons." Reasons are unique, and distinct from "mere causes," in two ways (Malle 1999; Malle et al. 2000).

First, reasons are considered to emerge from a unique reasoning process, moving from beliefs and desires to intentions and intentional actions (which classic BDI architectures have already worked to implement). This process confers a form of rationality to the action: Considering the particular agent's beliefs and desires, it makes sense for this agent to act this way; it is rational to do so. Second, reasons carry a form of subjectivity. People say, "His reasons make no sense"; or they ask, "What were her reasons?" (they don't ask, "What were his causes for tripping?") When citing reasons, explainers select what they deem the critical steps in this particular agent's reasoning process that led up to the agent's intention or action. Reason explanations are therefore akin to acts of perspective taking: capturing the agent's subjective view of what made that action sensible. This becomes clearest when the explanation cites a false belief, which explains the action truly from the agent's subjective perspective, rather than in terms of objective reality ("Why is she carrying around the umbrella in bright sunshine?"—"She thought it would rain."). By contrast, people offer cause explanations in an attempt to cite objective causal relations.

The two assumptions of rationality and subjectivity, when applied to AIS, demand certain capacities from such systems, ones that are by no means unusual. For one thing, people gladly grant robots rationality (Malle and Thapa Magar 2017), and rational thought is a hallmark of AI and robots (Parkes and Wellman 2015). For another, the assumption of subjectivity does not involve any mysterious self-awareness; it implies merely that the explanation of an agent's intentional action clarifies in what way the given action furthers this agent's desires in light of this agent's beliefs (even false ones). As long as agents have beliefs and desires different from those of other agents and can act on their own beliefs and desires, explanations of AIS actions can be "subjective" in the relevant way.

When Do People Seek Behavior Explanations?

Before we move to the challenges of implementing the outlined framework of behavior explanation to the case of AIS we need to address one more question. When do people explain behavior, and what behaviors are they most eager to explain? In short, people try to explain any given behavior (in self or other) if either (a) they themselves wonder why the behavior occurred or (b) they expect that someone else wonders why the behavior occurred. Case (a) leads to "private explanations," and case (b) leads to "communicative explanations." However, in both cases a wondering why initiates the process. People wonder why an event occurs if three conditions are met (Malle and Knobe 1997b): (1) the person is aware of the event, (2) does not yet (feel they) understand the event, and (3) finds the potential explanation personally relevant. As a result of these conditions, evidence shows (Malle and Knobe 1997b), people are most interested in explaining other people's intentional and publicly observable events. At the same time, we know from numerous studies that people explain (and expect others to explain) those intentional observable behaviors by referring to unobservable mental states: beliefs, desires, and intentions (for a review, see (Malle 2011)). Thus, explainable AIS, too, must be ready to explain their intentional actions (planned or already performed) by reference to their beliefs, desires, and intentions; and they must be ready to do so when humans they interact with wonder why the robot performed a certain behavior.

Challenges in Implementing Behavior Explanations in Autonomous Intelligent Systems

As AIS become more complex and ubiquitous, people will increasingly demand that they provide explanations for their actions. When designing such explainable AIS, what kinds of challenges will we face? Below we consider four design challenges and discuss potential paths to meet these challenges.

The Language of Explanation

The behavior explanations people expect of AIS will have to be in the language familiar to people as communicators and audiences of ordinary explanations. Suppose two people are having a conversation and one of them suddenly gets up, in the middle of the other's sentence, and is about to leave the room. It clearly wouldn't be an acceptable explanation to say, "I do this in order to maximize my rewards" or "Because it is the next step in my optimal policy." The interrupted conversation partner is owed a polite "I am sorry" followed by an explanation with a specific goal, such as "I need to go to bathroom" or a relevant belief, such as "I feel really sick." People who interact with AIS will expect such explanations from the system itself (not from a manual or a help line), and in natural language. When a healthcare robot declines the care recipient's request for an increase in pain medication, it might say, "I am not allowed to change your pain medication without your doctor's consent, and I have not yet been able to reach her." When a hotel guest enters her room, and finds a robot circling the bed, the robot might say, "I hope I am not disturbing you; my duty is to tidy up your room." Indeed, studies show that people prefer explanations given by AIS that are compatible with the intentional stance (Harbers, van den Bosch, and Meyer 2009)—that is, explanations referring to beliefs, desires and other mental states that motivated their decisions. Explainable AIS should therefore have the ability to clarify their actions by offering the reasons for those actions—the goals, beliefs, duties, and so

forth, that motivate and justify the actions (Langley et al. 2017; Broekens et al. 2010).

Distinct Classes of Behavior, Distinct Explanations

Our analysis of the conceptual framework of behavior explanation suggests that two cognitive capacities must be present in any AIS that explains its behavior the way people expect it to. First, the AIS must be able to distinguish intentional from unintentional behaviors (at least in itself but, ideally, also in other agents). Second, the system must be able to explain each of these classes of behavior in the expected way —unintentional behaviors with (mere) causes, intentional behaviors with reasons. We can examine the prospects for each of these aims against the backdrop of BDI (Belief, Desire, Intention) agents, which are of course modeled after the folk-conceptual framework of mind and action (Adam and Gaudou 2016).

BDI agents have several goals (desires), and they select behaviors based on whether they satisfy a goal or sub-goal, given what the agent believes about the behavior and the current state of the world (Broekens et al. 2010). Whenever the agent undergoes such reasoning steps it can register that it attempted an intentional action. However, the agent needs more than that to distinguish intentional from unintentional behavior, because some attempts for intentional action do not succeed. The agent therefore needs to have a predictive forward model of the likely outcomes in the world if the intentional action succeeds; outcomes that don't match the prediction are unintentional (either failures or undesired side-effects). In addition, the agent will need to track unintentional events that it causes in the absence of BDI planning (e.g., another agent pushes the AIS, and the AIS, while falling, damages property). These basic rules will need to be refined to account for complex, multi-layered actions and for multi-agent caused outcomes, but in principle we see no obstacle for AIS to correctly classify their own behaviors in terms of the human concept of intentionality.

Once they can distinguish unintentional and intentional behaviors, AIS will have to explain these two classes of behavior in distinct ways. Explanations of unintentional behaviors (including errors or side-effects) will refer to individual or sets of causes. Such causes may be formalized using causal models (Halpern and Hitchcock 2013; Sloman 2005), but they must still be translated into verbal representations so the audience readily comprehends them. This will require expressive vocabulary, including concepts of cause, omission, allowing, counterfactuals, and the like. By contrast, because BDI agents base their intentional actions on reasoning over desires and beliefs, they should be able to track these reasoning steps and therefore know in principle why they performed a particular intentional action (Broekens et al. 2010).

Selecting Relevant Explanations

Even though the ability to track one's reasoning steps is necessary and useful, it does not generate good behavior explanations. People do not want to hear a complete account of all the beliefs, goals, subgoals, or rejected actions that a system tracked. In conversations with humans as well as AIS, people prefer shorter explanations over longer ones, but also detailed ones over abstract ones (Harbers, van den Bosch, and Meyer 2009).This suggests that people keep some kind of relevance criterion in mind when they select and evaluate explanations. In the psychological literature, this is sometimes called the "causal selection problem"—the difficulty of selecting a small number of causes/reasons that sufficiently explain the event in question (Hesslow 1988; Hilton 2007).

How do people solve this problem? They determine what exact question the audience is interested in (McClure and Hilton 1998); they take into account what their audience member already knows (Slugoski et al. 1993); and they offer elements of explanations that build bridges between presumed knowledge and novel information (Korman and Malle 2016). In short, they offer explanations that generate coherence in a knowledge structure of old and new information (Thagard 1989). Although creating such coherence in current AIS is challenging, the problem may be solvable if people's conceptual framework of behavior explanation can be formalized and the missing elements in the structure, or the most informative additions to the structure, can be identified. We have seen earlier that reason explanations are based on structural relations between desires/goals, beliefs, and resulting intentional actions. These relations are often characterized as "practical reasoning arguments" (Walton 2015) and as conforming to a rationality principle (Malle 1999). There is reason for optimism that these structural relations can be formalized (Atkinson, Bench-Capon, and McBurney 2006).

From Structure to Content

Another major challenge, however, is the fact that structural relations alone do not suffice; reason explanations also incorporate contentful information. Knowing that an agent had *some* desire that he believed could be fulfilled by *some* action does not satisfactorily explain that action; *what* that desire was is critical for understanding the action. The proper relationship among contents of beliefs, desires, and actions requires there to be a large knowledge structure of associative, social, and causal relations (the dreaded "common sense"). Consider the following two explanations, one consistent with "common sense," the other one inconsistent:

(3) She took off her sunglasses because it was dark in the room and she wanted to better see what was going on.

(4) She took off her sunglasses because it was bright in the room and she wanted to better hear what was going on.

The reader is invited to write out the number of background beliefs that are necessary to deliver a complete deductive argument in place of explanation (3). That number will be large but likely manageable. And if AIS are deployed in reasonably constrained contexts, we may be able to provide many of the background beliefs necessary to comprehend and construct such explanations and let the system learn many additional ones.

It is clear that other challenges await, not the least of which are demands on natural language processing and pragmatic understanding under asymmetric knowledge conditions (e.g., when the explainer knows much more than the audience). Some technological capacities to support action explanations by AIS already exist, such as plan monitoring, conversational question answering, and storing and accessing of episodic memories (Langley et al. 2017). Therefore, we are confident that formalizing ordinary behavior explanations is feasible, and implementing them amenable in AIS is technically possible. At the same time, we believe that failing to take into account people's conceptual framework of mind and behavior would create insurmountable obstacles to the success of such implementations.

Conclusion and Outlook

When AIS explain their actions, people are more likely to form correct mental models of such systems, which enables them to calibrate their trust in these systems. We propose that AIS regarded by their users as intentional agents must generate explanations within the bounds of the conceptual and linguistic framework of human behavior explanation. In this paper, we have explored how this framework can be applied to AIS in order to make such systems "explainable" and "transparent." Challenges loom, such as designing the proper language of explanation, equipping AIS with the cognitive capacities to distinguish intentional from unintentional behaviors, and the ability to select the most relevant explanation out of a potentially very large set of causes or reasons. However, we see no principled obstacles to the design of AIS that meet these challenges.

Because our proposed approach to developing explainable AIS builds up the framework of human behavior explanation, human judgments should also play a central role in the evaluation of AIS to become more transparent. Such evaluations would probe people's perceptions of how natural, understandable, and appropriate a given explanation is, or people's level of trust they put in explainable (vs. opaque) AIS, including their willingness to interact with the system in the future. Objective evaluations would assess people's ability to calibrate their trust in an AIS as a function of the quality of its explanations and to form accurate mental models of an explainable AIS's capacities and expected actions. Additionally, future work will need to broaden the context of explanation and recognize that explanations occur primarily in interpersonal settings, where, for example, people indicate to the AIS their confusion or gratefully acknowledge their newly gained understanding. If people reach such understanding, they will not only trust the AIS's corresponding future action but also teach the agent if a given action was inappropriate or particularly laudable. This feedback will adjust the AIS's beliefs or rearrange its goal priorities, thus making explainable AI teachable AI.

Acknowledgements

This research was supported by a Rubicon grant 446-16-007 from the Netherlands Organization for Scientific Research (NWO) and by a grant from the Office of Naval Research (ONR), N00014-16-1-2278. The opinions expressed here are our own and do not necessarily reflect the views of NWO or ONR.

References

Adam, Carole, and Benoit Gaudou. 2016. "BDI Agents in Social Simulations: A Survey." *The Knowledge Engineering Review* 31 (03): 207–38.

Atkinson, Katie, Trevor Bench-Capon, and Peter McBurney. 2006. "Computational Representation of Practical Argument." *Synthese* 152 (2): 157–206.

Broekens, Joost, Maaike Harbers, Koen Hindriks, Karel van den Bosch, Catholijn Jonker, and John-Jules Meyer. 2010. "Do You Get It? User-Evaluated Explainable BDI Agents." In *Multiagent System Technologies*, edited by Jürgen Dix and Cees Witteveen, 28–39. Berlin, Heidelberg: Springer.

Bromberger, Sylvain. 1965. "An Approach to Explanation." In *Analytical Philosophy*, edited by R. Butler, 2nd series:72–105. Oxford: Basil Blackwell.

Bybee, Joan L, Revere D Perkins, and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Cleland-Huang, Jane, Orlena Gotel, and Andrea Zisman, eds. 2012. *Software and Systems Traceability*. London: Springer.

D'Andrade, R. G. 1987. "A Folk Model of the Mind." In *Cultural Models in Language and Thought*, edited by D. Holland and N. Quinn, 112–48. New York, NY: Cambridge University Press.

Dittrich, Winand H., and Stephen E. G. Lea. 1994. "Visual Perception of Intentional Motion." *Perception* 23 (3): 253–68.

Epley, Nicholas, Adam Waytz, and John T. Cacioppo. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." *Psychological Review* 114 (4): 864–86.

Fisher, Michael, Louise Dennis, and Matt Webster. 2013. "Verifying Autonomous Systems." *Communications of the ACM* 56 (9): 84–93.

Halpern, Joseph Y., and Christopher Hitchcock. 2013. “Compact Representations of Extended Causal Models.” *Cognitive Science* 37 (6): 986–1010.

Harbers, Maaike, Karel van den Bosch, and John-Jules Ch. Meyer. 2009. “A Study into Preferred Explanations of Virtual Agent Behavior.” In *Intelligent Virtual Agents*, edited by Zsófia Ruttkay, Michael Kipp, Anton Nijholt, and Hannes Högni Vilhjálmsson, 132–45. Berlin, Heidelberg: Springer.

Heider, Fritz. 1958. “The Naïve Analysis of Action.” In *The Psychology of Interpersonal Relations*, 79–124. New York: Wiley.

Hesslow, G. 1988. “The Problem of Causal Selection.” In *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*, edited by Denis J. Hilton, 11–32. Brighton, UK: Harvester Press.

Hilton, Denis J. 1990. “Conversational Processes and Causal Explanation.” *Psychological Bulletin* 107: 65–81.

———. 2007. “Causal Explanation: From Social Perception to Knowledge-Based Causal Attribution.” In *Social Psychology: Handbook of Basic Principles*, edited by Arie W. Kruglanski and E. Tory Higgins, 2nd ed., 232–53. New York, NY: Guilford Press.

Hoffman, Guy, Gurit E. Birnbaum, Keinan Vanunu, Omri Sass, and Harry T. Reis. 2014. “Robot Responsiveness to Human Disclosure Affects Social Impression and Appeal.” In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction (HRI ’14)*, 1–8. New York: ACM Press.

Horgan, T., and J. Woodward. 1985. “Folk Psychology Is Here to Stay.” *Philosophical Review* 94: 197–226.

Huang, Chien-Ming, and Bilge Mutlu. 2012. “Robot Behavior Toolkit: Generating Effective Social Behaviors for Robots.” In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI ’12)*, 25–32. New York: ACM.

Jackendoff, Ray, and Peter W. Culicover. 2003. “The Semantic Basis of Control in English.” *Language* 79 (3): 517–56.

Johnson, Susan C. 2000. “The Recognition of Mentalistic Agents in Infancy.” *Trends in Cognitive Sciences* 4 (1): 22–28.

Kiesler, Sara. 2005. “Fostering Common Ground in Human-Robot Interaction.” In *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication, ROMAN 2005*, 729–34. IEEE.

Kim, Taemie, and Pamela Hinds. 2006. “Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction.” In *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)*, 80–85. Hatfield, UK.

Korman, Joanna, and Bertram F. Malle. 2016. “Grasping for Traits or Reasons? How People Grapple with Puzzling Social Behaviors.” *Personality and Social Psychology Bulletin* 42 (11): 1451–65.

Langley, Pat, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. “Explainable Agency for Intelligent Autonomous Systems.” In *Proceedings of the Twenty-Ninth Innovative Applications of Artificial Intelligence Conference (IAAI-16)*, 4762–4764. AAAI Press.

Lee, Min Kyung, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski. 2010. “Gracefully Mitigating Breakdowns in Robotic Services.” In *Proceedings of the 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 203–10. IEEE.

Levin, Dan, Caroline Harriott, Natalie A. Paul, Tao Zheng, and Julie A. Adams. 2013. “Cognitive Dissonance as a Measure of Reactions to Human-Robot Interaction.” *Journal of Human-Robot Interaction* 2 (3): 1–17.

Lomas, M., R. Chevalier, E.V. Cross, R.C. Garrett, J. Hoare, and M. Kopack. 2012. “Explaining Robot Actions.” In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 187–88. Boston, MA.

Malle, Bertram F. 1999. “How People Explain Behavior: A New Theoretical Framework.” *Personality and Social Psychology Review* 3 (1): 23–48.

———. 2004. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. Cambridge, MA: MIT Press.

———. 2005. “Folk Theory of Mind: Conceptual Foundations of Human Social Cognition.” In *The New Unconscious*, edited by Ran R. Hassin, James S. Uleman, and John A. Bargh, 225–55. New York, NY: Oxford University Press.

———. 2011. “Time to Give up the Dogmas of Attribution: A New Theory of Behavior Explanation.” In *Advances of Experimental Social Psychology*, edited by Mark P. Zanna and James M. Olson, 44:297–352. San Diego, CA: Academic Press.

Malle, Bertram F., and Joshua Knobe. 1997a. “The Folk Concept of Intentionality.” *Journal of Experimental Social Psychology* 33 (2): 101–21.

———. 1997b. “Which Behaviors Do People Explain? A Basic Actor-Observer Asymmetry.” *Journal of Personality and Social Psychology* 72 (2): 288–304.

———. 2001. “The Distinction between Desire and Intention: A Folk-Conceptual Analysis.” In *Intentions and Intentionality: Foundations of Social Cognition.*, edited by Bertram F. Malle, Louis J. Moses, and Dare A. Baldwin, 45–67. Cambridge, MA: MIT Press.

Malle, Bertram F., Joshua Knobe, Matthew J. O’Laughlin, Gale E. Pearce, and Sarah E. Nelson. 2000. “Conceptual Structure and Social Functions of Behavior Explanations: Beyond Person-Situation Attributions.” *Journal of Personality and Social Psychology* 79 (3): 309–26.

Malle, Bertram F., Louis J. Moses, and Dare A. Baldwin. 2001. *Intentions and Intentionality: Foundations of Social Cognition*. Cambridge, MA: MIT Press.

Malle, Bertram F., and Stuti Thapa Magar. 2017. “What Kind of Mind Do I Want in My Robot? Developing a Measure of Desired Mental Capacities in Social Robots.” In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 195–196. HRI ’17. New York, NY, USA: ACM.

McClure, John, and Denis J. Hilton. 1998. “Are Goals or Preconditions Better Explanations? It Depends on the Question.” *European Journal of Social Psychology* 28 (6): 897–911.

Monroe, Andrew E., Kyle D. Dillon, and Bertram F. Malle. 2014. “Bringing Free Will down to Earth: People’s Psychological Concept of Free Will and Its Role in Moral Judgment.” *Consciousness and Cognition* 27 (July): 100–108.

Mueller, Erik T. 2016. *Transparent Computers: Designing Understandable Intelligent Systems*. CreateSpace Independent Publishing Platform.

Ohtsubo, Y. 2007. “Perceived Intentionality Intensifies Blameworthiness of Negative Behaviors: Blame-Praise Asymmetry in

Intensification Effect." *Japanese Psychological Research* 49 (2): 100–110. doi:10.1111/j.1468-5884.2007.00337.x.

Parkes, D. C., and M. P. Wellman. 2015. "Economic Reasoning and Artificial Intelligence." *Science* 349 (6245): 267–72.

Phillips, Ann T., Henry M. Wellman, and Elizabeth S. Spelke. 2002. "Infants' Ability to Connect Gaze and Emotional Expression to Intentional Action." *Cognition* 85 (1): 53.

Powers, A., Adam D. I. Kramer, S. Lim, J. Kuo, Sau-lai Lee, and S. Kiesler. 2005. "Eliciting Information from People with a Gendered Humanoid Robot." In *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication, 2005. ROMAN 2005*, 158–63.

Premack, David. 1990. "The Infant's Theory of Self-Propelled Objects." *Cognition* 36 (1): 1–16.

Premack, David, and Guy Woodruff. 1978. "Does the Chimpanzee Have a Theory of Mind?" *Behavioral and Brain Sciences* 1 (04): 515–26.

Robinette, Paul, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. 2016. "Overtrust of Robots in Emergency Evacuation Scenarios." In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction, HRI '16*, 101–108. Piscataway, NJ: IEEE Press.

Salem, Maha, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. "Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust." In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI '15*, 141–148. New York: ACM.

Searle, J. R. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Sloman, Steven A. 2005. *Causal Models: How People Think about the World and Its Alternatives*. New York: Oxford University Press.

Slugoski, Ben R., Mansur Lalljee, Roger Lamb, and Gerald P. Ginsburg. 1993. "Attribution in Conversational Context: Effect of Mutual Knowledge on Explanation-Giving." *European Journal of Social Psychology* 23 (3): 219–38.

Thagard, Paul. 1989. "Explanatory Coherence." *Behavioral and Brain Sciences* 12 (3): 435–502.

Theodorou, Andreas, Robert H. Wortham, and Joanna J. Bryson. 2016. "Why Is My Robot Behaving like That? Designing Transparency for Real Time Inspection of Autonomous Robots." In *AISB Workshop on Principles of Robotics*, 2016, University of Sheffield.

Triebel, Rudolph, Kai Arras, Rachid Alami, Lucas Beyer, Stefan Breuers, Raja Chatila, Mohamed Chetouani, et al. 2016. "SPENCER: A Socially Aware Service Robot for Passenger Guidance and Help in Busy Airports." In *Field and Service Robotics*, 607–22. Springer, Cham.

Tullio, Joe, Anind K. Dey, Jason Chalecki, and James Fogarty. 2007. "How It Works: A Field Study of Non-Technical Users Interacting with an Intelligent System." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07*, 31–40. New York: ACM.

Voiklis, John, Boyoung Kim, Corey Cusimano, and Bertram F. Malle. 2016. "Moral Judgments of Human vs. Robot Agents." In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 486–91. IEEE.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. "Transparent, Explainable, and Accountable AI for Robotics." *Science Robotics* 2 (6): eaan6080. doi:10.1126/scirobotics.aan6080.

Walton, Douglas N. 2015. *Practical Reasoning*. Vol. 2. Wiley Encyclopedia of Management. John Wiley & Sons.

Wang, Ning, David V. Pynadath, and Susan G. Hill. 2016. "Trust Calibration within a Human-Robot Team: Comparing Automatically Generated Explanations." In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction, HRI '16*, 109–116. Piscataway, NJ: IEEE Press.

Williams, Tom, Priscilla Briggs, and Matthias Scheutz. 2015. "Covert Robot-Robot Communication: Human Perceptions and Implications for HRI." *Journal of Human-Robot Interaction* 4 (2): 23–49.

Wortham, Robert H., Andreas Theodorou, and Joanna J. Bryson. 2016a. "Robot Transparency, Trust and Utility." In AISB Workshop on Principles of Robotics, 2016, University of Sheffield.

———. 2016b. "What Does the Robot Think? Transparency as a Fundamental Design Requirement for Intelligent Systems." In *Proceedings of the 2016 IJCAI Workshop on Ethics for Artificial Intelligence, New York*.