Challenges and solutions for trustworthy explainable robots

Guglielmo Papagni, Sabine T. Koeszegi

Abstract

For robots to be accepted within society, non-expert users must deem them not only useful (and usable) but also trustworthy. Designing robots that can explain their decisions and actions in terms that everyone can understand is crucial to their trustworthiness and successful integration into our society. This paper, written as a part of a doctoral dissertation, draws from interdisciplinary research on social sciences and explainable robots (and AI) to address the set of challenges associated with making robots explainable and trustworthy. Particular attention is paid to non-expert users' perspectives within the context of everyday interactions. We claim that, as perfect explanations do not exist, their success in triggering understanding and fostering trust is determined by their plausibility. Furthermore, we maintain that plausible explanations are the result of contextual negotiations between the parties involved. As a result, this paper presents strategies formalized into a model for explanatory interactions to maximize users' understanding and support trust development.

Keywords

Explainable Robots, Trust, Non-Expert Users, Everyday Explanations

1 Introduction

Recently, the concept that AI and robots should be able to explain their inner workings, decisions, and actions has emerged in academic and societal discussions. Furthermore, as AI and robots permeate society at different levels, affecting people's everyday life, their decision-making processes should be understandable not only for machine learning and robotics experts but also for a broader audience of domain experts (i.e., practitioners from fields where AI technologies are applied) and non-expert end-users. Importantly, each of these categories of users has different demands in terms of explainability desiderata and goals, as their interests and knowledge of the technology may differ substantially. To this extent, it is crucial to understand and acknowledge the differences between different categories of users and, hence, what explainability entails in each context.

The category of domain experts is concerned with applications, such as military operations (e.g., robots used for mine detection and removal or rescue tasks), exploration (e.g., in space or the oceans), and medical purposes. This implies that most of the users will need to undergo some sort of special training to interact with the robots. While this does not guarantee that these users will become robotics experts, such a training allows for creating an adequate mental model of the robot that, in turn, supports users' understanding and trust calibration. In contrast, the category of non-expert users refers to users who have little to no previous experience with specific robotic technologies. It includes application contexts such as caregiving and education, recreational activities, and, perhaps

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_3

most crucially, interactions with robots 'in the wild' [Sabanovic et al. 2006]. Because there has been no previous interaction or any introduction, the level of uncertainty concerning robots is higher in these contexts. According to several definitions, uncertainties and perception of risk represent two elements that may jeopardize trust [Lee and See 2004; Andras et al. 2018; Luhmann 2018].

This paper addresses a set of challenges of making robots explainable and trustworthy, particularly for non-expert users and within the context of everyday interactions. The main reason for doing this project is those non-expert users represent the vast majority of the public, and many robots and other AI-based technologies are designed to interact with them daily. Furthermore, because of their lack of technical knowledge and agency to manipulate robotic technologies, non-expert users are the most vulnerable. In this context, explainability plays a crucial and multifaceted role. According to some studies, explanations that are properly tailored to the needs of non-expert users reduce perceived uncertainty and increase the understandability of robots. This, in turn, supports users with trust calibration toward robots and, consequently, robot acceptability in society [Lomas et al. 2012; Langley 2016; Langley et al. 2017; Sheh 2017b; Andras et al. 2018; Papagni and Koeszegi 2020, 2021b]. Therefore, designing robots that can explain their decisions and actions in terms that everyone can understand will aid in their successful integration into our society. Furthermore, while the interests and needs of specific groups of users might differ, an explanation that is understandable by users with no prior knowledge of robotic technologies should be understandable to more technologically accustomed ones.

One of the major problems in tailoring robot explanations to the needs of non-expert users is that explainability is frequently considered a data-driven rather than goal-driven characteristic [Sado et al. 2020]. Instead, we claim that the design of social robots should integrate inputs from various disciplines and focus on developing the capacity to communicate decisions in terms easily graspable by a broad audience. Another problem that requires more extensive investigation is that explanations are, by their very nature, incomplete approximations of the actual decision-making processes [Keil 2006; Rudin 2018; Wang 2019]. The lack of perfect explanations is even more problematic for robotics, given the standardized, algorithmic, and 'coordinate-based' modalities of information processing that are typical of robots [Lomas et al. 2012].

We approach these challenges with an interdisciplinary drive. Seeking and providing explanations is a form of everyday social communication, which has been extensively studied within disciplines, such as philosophy, sociology, and psychology [Hilton 1990; Miller 2019]. Combining findings from such disciplines with the need to integrate them into the design of robots and other artificial agents can be labeled as an 'interdisciplinary challenge' of explainability [Adadi and Ber-

rada 2018]. Specifically, this paper discusses the core elements of a recent model for explanatory interactions with artificial agents proposed by the authors of this paper (see figure 1). The remainder of this paper is organized as follows. Section 2 introduces the model and briefly analyzes its development and core elements. Concerning the standardization of explanations, Section 2 presents the concept of contextual, co-constructed plausibility as the most significant feature upon which explanations should be built. Section 3 addresses the timing of explanations, which represents a central element of the model, to answer the question of when explanations are mostly needed to support the trust calibration between users and robots. Furthermore, Section 3 briefly presents the results of a study conducted in the context of repeated interaction with a virtual agent, whose accuracy and explainability are manipulated. Section 4 discusses whether a robot's decision or action ought to be explained because of intentions and reasoning or other causes (e.g., natural or mechanical), as this aspect is critical for the structure of an explanation. Section 5 focuses on communication strategies to increase the explanations' understandability, particularly on the possibility of multimodal and interactive explanations, which is at the heart of the non-expert users' question. Section 6 concludes the paper by addressing limitations and outlining the direction for future work.

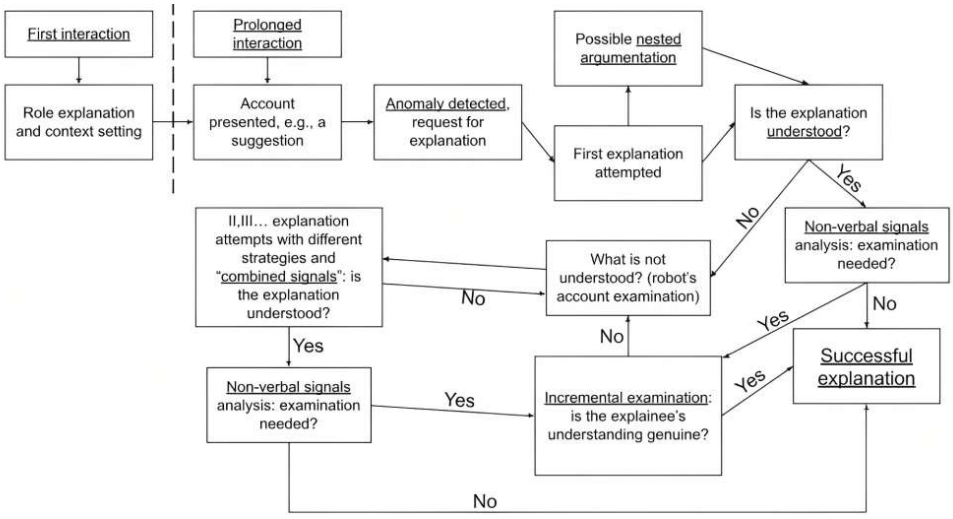

Figure 1 Explanatory Dialogue Model Adapted from [Papagni and Koeszegi 2021b]

2 Explanatory dialogue models

According to Berland, "literature in both the philosophy of science and psychology suggests that no single definition of explanation can account for the range of information that can satisfy a request for an explanation" [Berland and Reiser 2009, p. 27]. Accordingly, there is no single model to describe a perfect explanatory interaction. Furthermore, as previously stated, such models must be suitable for implementation in the algorithmic information-processing units of robots. The model presented in this paper aims to cope with this issue [Papagni and Koeszegi 2021b]. To do so, we analyzed existing models for explainable artificial intelligence (XAI), identified shortcomings, and developed solutions accordingly [Walton 2011; Madumal et al. 2018, 2019].

We identify two major limitations in Walton's, as well as Madumal's, Miller's, Sonenberg's, and Vetere's models. Sections 3 and 5 discuss more thoroughly each of these shortcomings. However, it is important to introduce them, as they both play central roles in the design and structure of our model. The first one concerns the timing of explanatory interactions and, more specifically, the notion that explanation requests are always promoted by an 'anomaly detection' [Walton 2011] or 'knowledge discrepancy' [Madumal et al. 2018, 2019]. This approach expresses the idea of explanations as isolated events, rather than as contextual instances. For instance, the models mentioned do not account for the fact that explanations concerning the robot's function in the specific interaction context are required at the beginning of an interaction with a robot, especially if this occurs 'in the wild'. This moment plays a role in how people build their mental model of the robot and should thus be considered part of the explanatory interaction.

The second shortcoming we identify is ensuring users' understanding of explanations. As previously noted, the inner workings of AI-based technologies are difficult to understand, even for expert users, let alone non-expert. If the robots' explanations are also not properly understood, the initial problem remains, since customers will still be unable to make sense of the robots' behavior. This argument also holds when applied to wrong explanations. How could an explanation be labeled as wrong if the content is not understood? Section 5 addresses these considerations more in detail.

2.1. Explanations' plausibility

Our model leverages on the principle of explanations' plausibility as the key criteria and ultimate goal [Papagni and Koeszegi 2020]. According to Karl Weick's 'sensemaking theory', sensemaking intended as a process, "is driven by plausibility rather than accuracy" [Weick et al. 2005, p. 415]. Building upon Peirce's work

on abductive reasoning, Wilkenfeld and Lombrozo rework Harman's concept of 'inference for the best explanation' [Harman 1965; Peirce 1997; Wilkenfeld and Lombrozo 2015]. Specifically, they postulate that the purpose of explainability should be to provide the best understanding of the causes of an event, rather than the most accurate explanation possible. This approach is consistent with Weick's idea that, to grasp the causes of an event, people seek plausible stories (i.e., that something 'might be') more than they seek true stories (i.e., that something 'actually is') [Peirce 1997; Miller 2019].

Malle argues that people seek explanations to find meanings and manage social interactions [Malle 2006]. According to Weick, the process of building meanings is the result of a collaborative effort involving the two parties (i.e., the explainer and explainee), as well as the context within which the interaction occurs [Weick et al. 2005]. In terms of explanatory interactions, there must be a knowledge transfer from the robotic explainer, who initially and 'asymmetrically' possesses the information that makes a specific explanation plausible, to the explainee, who must understand and agree that the explanation is plausible in that given context and for a specific event [Malle et al. 2007]. This does not necessarily imply that the explanation provided is the best in absolute terms, let alone the only one. The emphasis on all parties involved agreeing on the plausibility of an explanation implies the explainee's understanding of the explanation (i.e., it is unlikely for someone to find something plausible without understanding it in the first place). Furthermore, viewing plausibility as a collaborative and contextual achievement implies that the parties involved judge a given explanation as successful if it provides a satisfying account of an event's most likely causes.

Another advantage of adopting plausibility and abductive reasoning as core criteria of explainability is that there is no universally accepted principle for selecting a subset of causes upon which explanations are built. While certain qualities, such as internal coherence of an explanation and coherence of an explanation with prior beliefs, are generally considered desirable [Thagard 1989; Lombrozo 2007], the choice of other features is less obvious. For instance, some studies emphasize that explanations should be simple, whereas others consider complexity as the trademark of quality [Lombrozo 2007; Kulesza et al. 2013; Zemla et al. 2017]. If an explanation is only considered plausible when all the concerned parties agree, it follows that the most significant qualitative requirements for that situation are met. For an explanation to be (co-)considered plausible, the amount of information it conveys cannot be overwhelming or too scarce. Likewise, the explanation must be coherent with itself and with the prior beliefs of the concerned parties; it must not be too generic and vague, or complex, and so on. However, it could still be that an explanation will not be immediately considered plausible by all concerned parties. As plausibility is a quality that results from a negotiation,

multiple utterances may be required before all parties are satisfied. Section 5 discusses how this limitation can, at least in part, be dealt with.

2.2. Explainable robots, plausible robots

From these last considerations, plausibility is not a property that can be predefined once and for all. In other words, it is an aspect that is mostly determined by the context in which an interaction unfolds, the actors involved, and their specific interests. For instance, a possible application for social robots is assisting library customers. Among other tasks, such robots may suggest new readings to the customers, who may want to know the reasons for a specific recommendation before deciding. In a similar case, if the timing is not an issue, the robot may explain in detail how it arrived at that recommendation, by demonstrating how features, such as the customer's record of books requested in the past or feedback and reviews left by other users with similar preferences, weighed in the decision-making process [Ramos-Garijo et al. 2003; Mikawa et al. 2009; Sreejith et al. 2015]. Once these criteria have been presented by the robot, the customer may eventually agree (or disagree) with the explanation's plausibility and act accordingly.

However, in different situations, other features would likely be more relevant to show an explanation's plausibility. For instance, when the timing is an issue (e.g., during a rescue operation [Murphy 2004]), people may want robots to provide simple and concise explanations while not sparing vital information, particularly if the consequences of a wrong decision are potentially disastrous. In conclusion, what plausibility entails cannot (and probably should not) fall under an unambiguous, umbrella definition. The reason for this is that whether an explanation is plausible or not should be negotiated between the concerned parties, in a specific context.

3 Explanations' timing

This section focuses on the timing of explanations, a critical aspect that previous models have ignored, at least partially. Both models identify the start of an explanatory interaction in a 'knowledge discrepancy' or 'anomaly detection' [Walton 2011; Madumal et al. 2019]. Even though these models envision back-and-forth explanatory interactions with artificial agents, the type of approach they symbolize is one that ideally regards explanations as isolated instances. In contrast, we support Weick's view that meanings are co-constructed in the interplay between the concerned actors and the context, as we explain in the following paragraphs.

3.1. Initial explanations and trust formation

People provide explanations according to their mental model of the person with whom they are interacting in terms of the level of expertise and 'technicality' [Cawsey 1993]. In principle, this process is reliable because the parties involved in an explanatory interaction often share some knowledge about the topic being discussed. However, when it comes to robots, this can be problematic. When robots are employed in semi-controlled environments (e.g., in elderly care facilities or educational contexts), the researchers involved introduce them to users. To help users become acquainted with the robots, the researchers explain what the robots can and cannot do and support users in establishing an adequate initial mental model of the robots.

However, social robots are ultimately supposed to operate also 'in the wild' in everyday situations (e.g., at shopping malls and libraries) where people will mostly have little to no experience with robots and interactions will be limited in time. To this extent, initial trust depends on both personal attitude toward technology and 'institutional cues' [Siau and Wang 2018; Andras et al. 2018]. The former is a consequence of the combination of several factors, such as cultural background, demographics, and personality traits [Morris and Venkatesh 2000; Chien et al. 2016], and it can result in an equally wide range of dispositions toward new technologies, which are not necessarily mediated by accumulated experience with such technologies. These range from high expectations and over-trust [Dzindolet et al. 2003; De Visser et al. 2020], to skepticism and even forms of 'technophobia' [Kerschner and Ehlers 2016].

The notion that trust partially depends on 'institutional cues' refers to the role played by 'third parties', such as private companies, developers working for them, national and international institutions, and experts and regulatory bodies. Leveraging on their reliability and reputation, such entities play a 'proxy' role in determining how people perceive and trust new technologies. Specifically, this process is based on the assumption that the entities introducing new technologies act in accordance with values, such as integrity and benevolence, that define moral trust [Elia 2009; Lankton et al. 2015; Sood 2018]. Researchers have expressed concerns about the transparency, responsibility, and accountability of such 'third parties'. As for end-users initial trust in robots and AI, it is crucial to emphasize the importance of the adequate distribution of responsibilities (to, e.g., ensure technology transparency) among the stakeholders [Elia 2009; O'Leary 2019].

Based primarily on 'institutional cues' and individual attitude, initial trust can be very high or low irrespective of robots' actual performance concerning their purposes (i.e., not calibrated). For this reason, we emphasize the importance of the initial explanations. When robots have not yet proved to be reliable and

benevolent (e.g., on behalf of their makers), initial explanations may substitute the missing previous interactions, support the establishment of adequate mental models, and guide users toward placing calibrated trust in robots [Andras et al. 2018; Fossa 2019].

We agree with Cawsey that, in the event of a first-time interaction, robots should treat users 'as novices' which implies that robots should not assume anything about what users know. Accordingly, the robots' mental models of the users should only evolve and update as an interaction develops [Cawsey 1993]. According to Weick's argument, by adopting this approach, meanings and knowledge are lifted from the private and implicit sphere and made public and explicit [Weick et al. 2005]. Interestingly, Walton notes that "to grasp the anomaly, you have to be aware of the common knowledge" [Walton 2011, p. 365] and that "the system has to know what the user knows, to fill in the gaps" [Walton 2011, p. 365]. This appears to contradict the idea that explanation requests are triggered by the detection of an anomaly in one's account. However, how could a robot know what the user knows? Likewise, how can a user detect an anomaly in a robot's behavior if the user has no prior knowledge of what the robot should or should not do? For this reason, our model proposes that robots should provide initial explanations that contain basic information, such as what role and purpose the robot have and what it can and cannot do (see top left part of Figure 1). By so doing, robots could proactively establish the interaction context and support users in developing an adequate mental model. Additionally, once users are informed, the basic notions about the robot become shared knowledge and the robot can update its mental model of the user accordingly.

3.2. Unexpected events and trust restoration

According to the literature, the other moment in an interaction when people seek out explanations is when something unexpected or unpredictable happens [Andras et al. 2018; Miller 2019]. In other words, once users establish a mental model of a robot based on prior interactions, they will expect the robot to perform actions within a certain range of possibilities. Within this range, the robot's reliability will be progressively determined based on its performance and accuracy. As a robot regularly demonstrates reliability and trustworthiness, users may consolidate their positive mental model of it, so that explanations become superfluous if not even damaging [Doshi-Velez and Kim 2017]. However, a robot may still act unexpectedly or unpredictably. Such events, which do not fit into the established mental model, are also recorded in the interactions. This is also what the models by [Walton 2011; Madumal et al. 2018, 2019] label as 'knowledge discrepancy' or 'anomaly'. In similar situations, users' understanding of the robot's behavior

is challenged. Furthermore, several researchers have found that if users do not understand why a robot is behaving in a certain way, their acceptance and trust in the robot are probably weakened [Lomas et al. 2012; De Graaf and Malle 2017; de Graaf et al. 2018; Miller 2019]. While this is particularly the case when unexpected robot actions turn out to be mistakes [Elangovan et al. 2007; Robinette et al. 2017] even if a robot behaves according to its internal planning, if this is not obvious to users, it is important that they still make sense of why the robot is acting that way [Andras et al. 2018].

In an aging society, social robots are meant to be deployed in elderly-care facilities with assisting duties. IBM's MERA is one such robot being developed, on top of SoftBank Robotics' Pepper platform, for similar purposes. It can monitor people's pulse and breathing functions, among other things [Martinez-Martin and del Pobil 2018; Venkatesh 2019]. For instance, if the robot detects any anomalies in these parameters before the person is consciously aware of it, it may suggest the assisted person rest. Such an event may be perceived as an anomaly, prompting the assisted person to request an explanation, which would likely elucidate the reasons behind the suggestion and show that, while these reasons were not obvious at a first glance, they still make the robot's suggestion plausible.

Hence, whether it is to prevent the loss of trust, or restore it after a mistake, robots must provide reasons for their actions through explanations. Other trust restoration strategies, such as denial, apologies, compensation, and relationship restructuring, exist and can be implemented among robots' functions [Quinn et al. 2017; Lewicki and Brinsfield 2017]. However, unlike these strategies, explainability offers two main advantages. On the one hand, as we discussed in the previous paragraph, explanations support trust not only in the case of a violation but also in building it at the start of an interaction. On the other hand, explanations provide useful insights into the causes of an unexpected event or mistake. We previously noted that explanations may not be strictly necessary in case of repeated successful interactions with a robot. For instance, when "there are no significant consequences for unacceptable results" [Doshi-Velez and Kim 2017, p.3] or when a problem has been thoroughly researched and validated in real-world scenarios, explanations could become superfluous. However, even after multiple interactions, specific users may be unaware that a certain problem has been previously studied and that a robot's decision is based on real-world-validated data. Therefore, in principle, robots should always be able to explain themselves whenever users ask.

To examine some claims discussed in the previous sections, an empirical study was conducted. Participants were required to interact with a personalized virtual learning assistant seven times. The goal of the assistant was to provide participants with recommendations on what chunks of text to focus on (out of

larger portions), for them to prepare for quizzes. The system's explainability and accuracy were modified throughout the study.

Among the main findings, we observed that, contrary to expectations, initial explanations about the system's functionality did not increase initial trust. Simultaneously, the assistant's wrong recommendation affected participants' trust negatively, as it was perceived as a trust breach. However, qualitative data reveal that participants tended to be quite tolerant toward imperfect AI-based systems, as these systems are not expected to always function perfectly. Additionally, the qualitative data suggest that the researchers' 'hidden authority' has a favorable impact on the system's trustworthiness. Perhaps more importantly, trust restoration was significantly faster when the system provided an explanation following the wrong recommendation, rather than not. Specifically, explanations were the most effective as a trust-restoration strategy with risk-averse participants. Furthermore, explanations aided trust recovery, even if the participants did not always access them. Our qualitative analysis revealed how this may be explained, at least in part, by the fact that the very availability of explanations suggests a more transparent and trustworthy system.

4 Explainable robots and the intentional framework

Another element is crucial in terms of the mental model of robots and explanation generation. It is about whether or not robots' explanations ought to reflect some form of intentionality (and other mental states) behind robots' behavior. This aspect of explanatory interactions is part of a broader ongoing discussion between the human-robot and human-computer interaction (respectively, HRI and HCI) communities. While discussions on robots' 'mental states' have paced up recently, they have older roots that date back at least to Heider's and Simmel's work, as they demonstrated that people adopt a mentalistic framework to interpret even the movements of simple and schematic geometrical shapes [Heider and Simmel 1944]. Then, with Daniel Dennett's concept of the 'intentional stance', interest in the topic has spread. [Dennett 1988, 1989]. Dennett explained that people interact with certain technological artifacts (such as a chess-playing computer) as though they acted on human-like internal states, such as desires, beliefs, and intentions. According to Dennett, it would be too difficult to understand how such devices work solely by relying on one's knowledge of their intended purpose (i.e., the design stance), let alone the knowledge of natural laws (i.e., the physical stance) that ultimately govern everything [Dennett 1988, 1989, 1997]. Therefore, Dennett says, people adopt with computers and robots a mentalistic framework that is similar to that adopted with other people.

According to recent interpretations, the phenomenon is due to a 'primacy of the social mindset', which means that a mentalistic interpretative framework is always readily available because of people's social training and familiarity with it since childhood [Buckner et al. 2008; Looser and Wheatley 2010; Spunt et al. 2015; Papagni and Koeszegi 2021a]. Furthermore, as most people appear to lack a strategy for interacting specifically with sophisticated technologies, such as robots, a mentalistic approach eventually prevails. Attributing intentions to robots and other seemingly intelligent machines has some problematic aspects. For instance, researchers have proposed that in certain cases, the unconscious (and erroneous) adoption of a mentalistic framework may be the origin of the so-called 'uncanny valley' phenomenon [Bartneck et al. 2009; Mori et al. 2012]. Additionally, in certain situations, attempting to understand robots' behavior from a mentalistic perspective is not the best strategy, and users may have to forcefully adapt their mental model at the expense of cognitive resources [Wiese et al. 2017].

According to Weick's sensemaking framework, finding meanings in the social context of everyday life entails bringing order to the chaotic stream of both intentional behaviors and unintentional events. In terms of explanations, this translates to attributing either reasons, intentions, desires, and beliefs, or natural and mechanical causes. According to De Graaf and Malle, intentionality is a core concept that allows people to explain and understand others' behaviors [De Graaf and Malle 2017]. While the phenomenon has been thoroughly investigated in the human sciences, the concept of predicting and explaining robots' behavior using the intentionality framework is an open debate. According to Bossi, "people may treat robots as mechanistic artifacts or may consider them to be intentional agents. This might result in explaining robots' behavior as stemming from operations of the mind (intentional interpretation) or as a result of mechanistic design (mechanistic interpretation)" [Bossi et al. 2020, p. 1].

As we previously discussed, explanations are often sought after when users' mental models of robots are challenged by unpredictable events. This includes situations in which users cannot understand or explain robots' actions according to the mental model of robots they already possess. An implication of this interpretative gap is that whatever framework (i.e., intentional or mechanistic) users are adopting at the time of the unexpected occurrence, their trust in the framework's prediction-making power might decrease. In other words, when something unexpected happens, users may be unable to provide themselves with reasons or causes and, hence, ask the robot with whom they are interacting for an explanation. Some cases will force a complete perspective (i.e., framework) switch, while others will not. Importantly, according to De Graaf and Malle, robots "must be able to distinguish intentional from unintentional behaviors" and they "must be able to explain each of these classes of behavior in the expected way – unin-

tentional behaviors with (mere) causes, intentional behaviors with reasons" [De Graaf and Malle 2017, p. 19].

For instance, We previously mentioned, referring to elderly people's assistance, the possibility of the robot advising the assisted person takes a rest. The latter may not immediately grasp the reason for the recommendation, as they are unaware of what the robot knows. This includes not knowing whether the recommendation is genuine (i.e., based on the intention to assist the person) or based on a wrong premise (e.g., a malfunction). Assuming that the robot has been useful and has acted in the best interest of the user up to that point, the user may be struggling to make sense of the recommendation within the same (i.e., intentional) framework and may request an explanation. Within an intentional framework, the robot's explanation that its sensors have observed increased heart rate and heavy breathing would still make sense, as it would show the robot's intention to assist the user. A similar explanation emphasizes that the user was merely unaware of the robot's actual decision-making process. Accordingly, this implies that not every unpredictable behavior is the result of robots' malfunctions or internal errors, which are more likely to be detected (e.g., if the robot suddenly stops performing its tasks), and require users' to switch framework.

Ultimately, it could still be that a robot provides an explanation that makes sense (i.e., sounds plausible) within the boundaries of the framework adopted by the users but is built upon wrong premises [Dunne et al. 2005; Walton 2011]. As will be discussed in Section 5, when dealing with the structure of explanatory interactions, the risk of wrong explanations going unnoticed motivates taking further measures. Based on the discussion in this section, we claim that robots must be designed to support users, by means of explanations, in adopting the most appropriate interaction framework. This is especially the case for the early stages of extensive adoption of robotics in everyday contexts. Indeed, these times are most characterized by uncertainty in terms of both the adoption of and narratives built around these technologies. Furthermore, whenever necessary, robots should support the transition from one interpretative framework to another. We have previously discussed how the plausibility of explanations must be considered a contextual joint achievement. What framework is most adequate for understanding an event is a contextual feature that must be treated as such. Hence, robots should communicate explicitly and clearly, to the greatest extent feasible, whether the event being explained involves unintentional causes (e.g., an internal failure or mistake or uncontrollable external forces) or intentional reasons. In the next section, we will discuss explanation communication strategies that maximize the chances of users' correct understanding and hence trust toward the robots.

5 Communicating explanations

Explanations are primarily forms of social communication [Hilton 1990]. Therefore, addressing how robots should deliver explanations is likely the most essential aspect of explanatory interactions. This section analyzes two features of explanation communication that constitute the core of our model. Specifically, we discuss our claims that to support users' understanding and trust calibration, robots should:

- Be able to use diversified means of communication.
- Provide users with the possibility to question explanations and ask for further insights.

Importantly, when it comes to explainable robots, the research on the effects of combining the two mentioned strategies while promising is still in its early stages [Abdul et al. 2018; Anjomshoae et al. 2019].

5.1. Multi-modal explanation

In human-human interactions, explanations are primarily communicated through natural language. Generally, they should follow communication norms, such as 'Grice's (four) maxims of conversation' [Grice 1975]. They refer to communicating only what is confidently believed to be accurate, avoiding overwhelming amounts of information without being scarce, relevant to the context (i.e., a 'good social explanation' [Hellström and Bensch 2018; Miller 2019]), avoiding obscurity and ambiguity and being brief and orderly in presenting the information. Grice's maxims are often mentioned in explainable robots and AI research because they provide an implementable solution that may improve explanation quality [Miller 2019; Papagni and Koeszegi 2021b]. Sheh provides further possibilities for modifying how explanations are communicated through natural language [Sheh 2017a]. According to the author, robots can modify the depth and type of explanation based on the needs of specific interaction instances and the availability of the robots' underlying AI models. The author observes, in reference to a scenario in which a robotic shopping mall assistant is questioned about its product recommendations, that in similar circumstances, social robots' explanations are expected to primarily satisfy users' curiosity and support further engagement. For this reason, the author continues, 'Post-Hoc' explanations at 'Attribute Only' or 'Attribute Use' depths may be appropriate for the purpose [Sheh 2017a]. While the former indicates explanations that are tailored solely to what the robot deems the most relevant features, the latter considers the implications (i.e., 'use') of each attribute's value. Therefore, if properly tailored, text-based explanations alone already provide various customization options.

However, when the explanations' goal is to maximize users' understanding and trust calibration toward a robot, it is important to note that natural language only covers a subset of feasible communication strategies. Explanations in the form of 'combined signals' [Engle 1998], also known as 'multi-modal' explanations represent a promising but under-explored research avenue. Anjomshoae, Najjar, Calvaresi, and Främling discussed six possible communication modalities [Anjomshoae et al. 2019]. Besides text-based natural language explanations, they identified the "visualization" (i.e., graphical) type as the second most common one. Logs, expressive motions, expressive lights, and speech complete the list. The notion behind multimodality is that, as technological devices, robots can convey information through complementary modalities, sometimes even better than humans can. For instance, with visual explanations being the second most common after text-based ones, many robots can display on frontal screens graphic information gathered by their sensors, and once processed, these environmental data may support text to convey more complete messages. In our previous example, the IBM's MERA robot explained to the assisted person that its recommendation to take a rest was based on factors, such as the unusually high pulse rate and heavy breathing. While a text-based explanation would likely suffice to convey the essential message, the explanation's quality could still improve if the robot would provide visualizations of the actual scans of normal and abnormal heart activity. While HRI research on multimodality and 'combined signals' is still in its early stages, an increasing number of studies have demonstrated that users can benefit from multimodal explanations. The HCI community has done most of the research in multimodal explanations so far. Most studies effectively combined verbal and visual information, showing how people preferred this format to 'uni-modal' ones [Huk Park et al. 2018; Kanehira et al. 2019].

Two considerations must be made. First, the availability of alternative communication strategies should not mean that robots must display all available information at once. Explanations should not exclude vital information, but simultaneously, they should also not overwhelm users with too much information. To this extent, researchers propose that, in certain cases, employing alternative single-handed modalities may be more beneficial to the users. For instance, referring to robots' reactive planning, Theodoru, Wortham, and Bryson suggest that since robots can take many decisions per second, graphical explanations are more efficient and direct than verbal ones [Theodorou et al. 2016]. Giving self-driving systems the ability to employ light signaling to communicate simple messages to pedestrians, such as that they can cross the street safely [Faas and Baumann 2019], is another example of how alternative modalities can suffice even when taken alone. In conclusion, while in certain cases alternative modalities may provide adequate information, text-based explanations are likely to remain prominent (possibly sup-

ported by other means) because the semantic richness that can be conveyed through natural language is difficult to match through other means alone.

Finally, multimodality should not be unidirectional or limited to the combination of text-based and graphic communication. Natural language processing and image recognition have improved significantly recently, allowing robots and virtual agents to provide progressively better answers to users' text- or image-based inputs. One further possibility is that robots can 'read' and 'express' signals other than graphic and natural language communications. For instance, research in other relevant areas of robotics, such as (reading and expressing) body motion [Han et al. 2012; McColl and Nejat 2014] or facial expressions and gaze [Fiore et al. 2013; Admoni and Scassellati 2017] shows that robots can process various signal typologies that can make communication with humans (included explanatory interactions) more flexible and inclusive.

5.2. Interactive explanations

Making explanations 'interactive' is another promising strategy to increase robots' explanations quality that requires further investigation, particularly in the field of social robots [Abdul et al. 2018; Papagni and Koeszegi 2021b]. This research is partly driven by the desire to achieve a higher degree of human likeness [Madumal et al. 2018, 2019]. Indeed, explanations in robotics are often treated as 'single-shot' communication acts, whereas in human-human interaction, they frequently occur in the form of dialogues with back-and-forth iterations. However, interactivity also represents a strategy to deal with what Keil identifies as people's attitude to overestimate their own understanding of explanations (i.e., the 'illusion of explanatory depth') [Keil 2003]. According to Keil, this phenomenon, which is related to studies from social psychology on the 'introspection illusion' [Pronin 2009], consists of wrongly assessing the quality of the information one retains after being provided an explanation. The next paragraphs discuss our claim that, among other advantages, im- plementing design features that support interactivity of robots' explanations helps mitigate this phenomenon.

A fundamental contribution to the user-friendliness of explanations' interactivity is that it allows the parties involved to seek further insights to better understand what is being explained, and it allows questioning of both parties' accounts. The implementation of 'nested argumentation dialogues' [Madumal et al. 2018, 2019] and an 'examination phase' into our model aims to primarily tackle this multifaceted aspect [Dunne et al. 2005; Walton 2006, 2011].

Introducing nested argumentation dialogues allows users to engage in multilayered explanations in which they can drift from one question to another in a

back-and-forth manner. This back-and-forth movement may concern the topic of the original question or may be about 'spin-off' discussions [Madumal et al. 2018, 2019]. Often in human-human interaction, such spin-off argumentation dialogues are nested on top of the original explanation to support explainees by improving their understanding. The model proposed by Walton does not account for nested argumentation because the author labels overlapping dialog as an illicit dialectical shift, implying that the previous question must be considered closed [Walton 2011]. However, to achieve interaction naturalness and support users' sensemaking, robots should be able to process nested dialogs as such, leaving users the choice to return to the original one. Hence, to increase the human-likeness of explanatory interactions, our model allows users to engage in nested argumentation dialogues that are both related and unrelated to the original question, as shown in the top right corner of Figure 1. However, introducing such internal loops is merely one interpretation of the concept of interactivity.

Explanations may appear logical at a first glance and yet be grounded upon incorrect premises [Walton 2011; Dunne et al. 2005; Lakkaraju and Bastani 2020]. Introducing a dialectical shift in the form of an 'examination phase' allows users to analyze the explainer's account for any inconsistencies and evaluate the quality of the explanation for potential errors [Dunne et al. 2005; Lamche et al. 2014]. To this extent, Kaur et al. highlight the propensity, even among HCI expert practitioners, to over-rely on interpretability tools' visual outputs in a study in which they analyze participants' reactions to different approaches to model interpretability (i.e., 'glass-box' and 'black-box') To address this issue, one of their suggestions is to adopt 'back-and-forth explanations' (i.e., interactive interpretability) [Kaur et al. 2020].

Another possible use for an examination phase is to test the explainee's understanding of an explanation, as suggested by Walton [Walton 2011]. Indeed, as previously stated, people are susceptible to the 'illusion of explanatory depth' and tend to overestimate their understanding of explanations [Keil 2003]. Section 1 also highlighted the connections between understanding robots (and robots' explanations) and calibrating trust in them. For these reasons, assessments of understanding quality are an important aspect of models for explanatory interactions. This is supposed to be done by questioning the explainee about the explanation, the causal connections to the event being explained, and so on. Nevertheless, testing users' understanding should not translate into an interrogation, as this may be perceived as aggressive and have overall counterproductive effects on the interaction [Walton 2011]. To this extent, the authors of the model described in [Madumal et al. 2018, 2019] assert that such an operation is uncommon in everyday human-human interactions. Instead, to keep the interaction as natural as possible, they consider the explainee's affirmation of effective under-

standing as a sufficient criterion to measure the quality of the explanation. While we agree that explanatory interactions should feel natural and smooth to users, rather than making them feel uncomfortable and jeopardizing future interactions, we also acknowledge a gap in the model from Madumal, Miller, Sonenberg, and Vetere in terms of evaluation strategies for the success of an explanation. Therefore, we deem an 'incremental approach' to be the most appropriate [Papagni and Koeszegi 2021b]. Alternatively, after a robot provides an explanation, it may ask users to pick among multiple options what they understood to be the right explanation. To this end, we claim that testing users' understanding must be contextually calibrated based on how much time and interest users are willing to invest. In other words, instead of being predetermined by the robot, questions concerning the explanation must be negotiated with users based on contextual affordances.

Finally, just as it occurs in human-human interaction, it is impossible to guarantee the success of explanations in terms of knowledge transfer and users' understanding. Despite robots' best attempts, there will be circumstances in which users do not grasp what is being explained to them. Future research on explainable robots should focus on how to minimize the likelihood of such events occurring by refining and testing solutions, such as the ones presented in this paper, and implementing alternative strategies to better prevent trust losses and restore trust after a violation.

6 Future work and conclusions

The presence of social robots in everyday life is becoming a reality. Their successful integration and acceptability into society depend not only on how useful they prove to be in terms of performance but also on how they explain their decisions to a broad audience of non-expert users. At the same time, this paper acknowledges that perfect explanations do not exist and that making robots explainable poses a multifaceted interdisciplinary challenge. To solve this problem, we proposed a model for explanatory interactions. This model considers important findings from social sciences as well as from research on explainable AI and robots and their affordances and availability in terms of explainability. Furthermore, as the key criterion to assess the quality of explanations, we proposed a notion of explanations' plausibility as a joint achievement, which presupposes the users' understanding of robots' explanations.

One of the main limitations is that the type of explanation a robot can provide depends on the availability of the underlying algorithms and the physical capabilities of individual robots. In other words, not all the features of our model may

be implemented in the behavioral programming of certain robots. Therefore, research should focus on how to broaden the scope of both AI models' explainability and robots' customization. Another limitation concerns the primarily conceptual nature of the work presented in this paper. This calls for follow-up experimental studies to test our claims and the feasibility of implementing the various features of our model. Such studies shall, for instance, focus on the long-term effects of explanations on trust formation and restoration. Likewise, the combination of multimodal and interactive strategies is a promising but understudied research avenue that may shed further light on users' reception of explainable robots in terms of both trust and understandability.

Bibliography

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems.* 1–18.

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explain- able Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63.

Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R. Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T. Powers, Neil Urquhart, and Simon Wells. 2018. Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems. *IEEE Technology and Society Magazine* 37, 4 (2018), 76–83. https://doi.org/10.1109/MTS.2018.2876107

Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. *Explainable Agents and Robots: Results from a Systematic Literature Review*. International Foundation for Autonomous Agents and Multiagent Systems. http://dl.acm.org/citation.cfm?id=3306127.3331806

Christoph Bartneck, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. My robotic doppelgänger-A critical look at the uncanny valley. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 269–276.

Leema Kuhn Berland and Brian J Reiser. 2009. Making sense of argumentation and explanation. *Science education* 93, 1 (2009), 26–55.

Francesco Bossi, Cesco Willemse, Jacopo Cavazza, Serena Marchesi, Vittorio Murino, and Agnieszka Wykowska. 2020. The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. *Science robotics* 5, 46 (2020).

Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. 2008. The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences* 1124 (2008), 1–38. https://doi.org/10.1196/annals.1440.011

Alison Cawsey. 1993. User modelling in interactive explanations. *User Modeling and User-Adapted Interaction* 3, 3 (1993), 221–247. https://doi.org/10.1007/BF01257890

Shih-Yi Chien, Katia Sycara, Jyi-Shane Liu, and Asiye Kumru. 2016. Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 841–845.

Maartje MA De Graaf and Bertram F Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series.*

Maartje MA de Graaf, Bertram F Malle, Anca Dragan, and Tom Ziemke. 2018. Explainable robotic systems. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* 387–388.

Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics* 12, 2 (2020), 459–478.

Daniel C Dennett. 1988. Précis of the intentional stance. *Behavioral and brain sciences* 11, 3 (1988), 495–505.

Daniel C Dennett. 1989. *The intentional stance.* MIT press.

Daniel C Dennett. 1997. True Believers: The Intentional Strategy and Why It works. *Mind Design* (1997), 57–79.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608 (2017).

Paul E Dunne, Sylvie Doutre, and Trevor Bench-Capon. 2005. Discovering inconsistency through examination dialogues. In *Proceedings of the 19th international joint conference on Artificial intelligence*. 1680–1681.

Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.

A R Elangovan, Werner Auer-Rizzi, and Erna Szabo. 2007. Why don't I trust you now? An attributional approach to erosion of trust. *Journal of Managerial Psychology* (2007). 22(1), 4–24. https://doi.org/10.1108/02683940710721910

John Elia. 2009. Transparency rights, technology, and trust. *Ethics and Information Technology* 11, 2 (2009), 145–153.

Randi A Engle. 1998. Not channels but composite signals: Speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. In *Proceedings of the twentieth annual conference of the cognitive science society.* 321–326.

Stefanie M Faas and Martin Baumann. 2019. Yielding light signal evaluation for self-driving vehicle and pedestrian interaction. In *International Conference on Human Systems Engineering and Design: Future Trends and Applications.* Springer, 189–194.

Stephen M Fiore, Travis J Wiltshire, Emilio JC Lobato, Florian G Jentsch, Wesley H Huang, and Benjamin Axelrod. 2013. Toward understanding social cues and signals in human–robot interaction: effects of robot gaze and proxemic behavior. *Frontiers in psychology* 4 (2013), 859.

Fabio Fossa. 2019. I don't trust you, you faker! On trust, reliance, and artificial agency. *Teoria*, 1(XXXIX) (2019), 63–80.

Herbert P Grice. 1975. Logic and conversation. In *Syntax and semantics. Vol. 3, Speech acts*, P. Cole und J. L. Morgan (Ed.). Brill, 41–58.

JingGuang Han, Nick Campbell, Kristiina Jokinen, and Graham Wilcock. 2012. Investigating the use of non-verbal cues in human-robot interaction with a Nao robot. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom).* IEEE, 679–683.

Gilbert H Harman. 1965. The inference to the best explanation. *The philosophical review* 74, 1 (1965), 88–95.

Fritz Heider and Marianne Simmel. 1944. An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57, 2 (1944), 243–259.

Thomas Hellström and Suna Bensch. 2018. Understandable robots-what, why, and how. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 110–123.

Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 8779–8788.

Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, and Tatsuya Harada. 2019. Multimodal explana- tions by predicting counterfactuality in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8594–8602.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–14.

Frank C Keil. 2003. Folkscience: Coarse interpretations of a complex reality. *Trends in cognitive sciences* 7, 8 (2003), 368–373.

Frank C Keil. 2006. Explanation and understanding. *Annual Review of Psychology.* 57 (2006), 227–254.

Christian Kerschner and Melf-Hinrich Ehlers. 2016. A framework of attitudes towards technology in theory and practice. *Ecological Economics* 126 (2016), 139–151.

Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 3–10.

Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

Béatrice Lamche, Ugur Adıgüzel, and Wolfgang Wörndl. 2014. Interactive explanations in mobile shopping recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*, Vol. 14.

Pat Langley. 2016. Explainable agency in human-robot interaction. In *AAAI Fall Symposium Series.*

Pat Langley, Ben Meadows, Mohan Sridharan, and Dongkyu Choi. 2017. Explainable agency for intelligent autonomous systems. In *Twenty-Ninth IAAI Conference*.

Nancy K Lankton, D Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems* 16, 10 (2015), 880-918.

John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

Roy J Lewicki and Chad Brinsfield. 2017. Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior* 4 (2017), 287–313.

Meghann Lomas, Robert Chevalier, Ernest Vincent Cross, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction.* 187–188.

Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology* 55, 3 (2007), 232–257.

Christine E Looser and Thalia Wheatley. 2010. The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological science* 21, 12 (2010), 1854–1862.

Niklas Luhmann. 2018. *Trust and power.* John Wiley & Sons.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. *arXiv preprint* arXiv:1903.02409 (2019).

Prashan Madumal, Tim Miller, Frank Vetere, and Liz Sonenberg. 2018. Towards a grounded dialog model for explainable artificial intelligence. *arXiv preprint* arXiv:1806.08055 (2018).

Bertram F Malle. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press.

Bertram F Malle, Joshua M Knobe, and Sarah E Nelson. 2007. Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of personality and social psychology* 93, 4 (2007), 491–514.

Ester Martinez-Martin and Angel P del Pobil. 2018. Personal robot assistants for elderly care: an overview. In *Personal assistants: Emerging computational technologies,* A Costa, V. Julian & P. Novaris (Ed.). (2018), 77–91. Springer International Publishing. https://doi.org/10.1007/978-3-319-62530-0_5

Derek McColl and Goldie Nejat. 2014. Recognizing emotional body language displayed by a human-like social robot. *International Journal of Social Robotics* 6, 2 (2014), 261–280.

Masahiko Mikawa, Masahiro Yoshikawa, Takeshi Tsujimura, and Kazuyo Tanaka. 2009. Librarian robot controlled by mathematical aim model. In *2009 ICCAS-SICE. IEEE*, 1200–1205.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

Michael G Morris and Viswanath Venkatesh. 2000. Age differences in technology adoption decisions: Implications for a changing work force. *Personnel psychology* 53, 2 (2000), 375–403.

Robin R Murphy. 2004. Trial by fire [rescue robots]. *IEEE Robotics & Automation Magazine* 11, 3 (2004), 50–61.

Daniel E O'Leary. 2019. GOOGLE'S Duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management* 26, 1 (2019), 46–53.

Guglielmo Papagni and Sabine Koeszegi. 2020. Interpretable Artificial Agents and Trust: Supporting a non-Expert Users Perspective. In *Culturally Sustainable Social Robotic*, M. Nørskov, J. Seibt, O. S. Quick (Eds.). (2020), IOS Press, Amsterdam. 653–662.

Guglielmo Papagni and Sabine Koeszegi. 2021a. A Pragmatic Approach to the Intentional Stance Semantic, Empirical and Ethical Considerations for the Design of Artificial Agents. *Minds and Machines* 31, 4 (2021), 505–534.

Guglielmo Papagni and Sabine Koeszegi. 2021b. Understandable and trustworthy explainable robots: A sensemaking perspective. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 13–30.

Charles Sanders Peirce. 1997. *Pragmatism as a principle and method of right thinking: The 1903 Harvard lectures on pragmatism.* SUNY Press.

Emily Pronin. 2009. The introspection illusion. *Advances in experimental social psychology* 41 (2009), 1–67.

Daniel B Quinn, Richard Pak, and Ewart J de Visser. 2017. Testing the efficacy of human-human trust repair strategies with machines. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 61. SAGE Publications Sage CA: Los Angeles, CA, 1794–1798.

Rafael Ramos-Garijo, Mario Prats, Pedro J Sanz, and Angel Pasqual Del Pobil. 2003. An autonomous assistant robot for book manipulation in a library. In *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, Vol. 4. IEEE, 3912–3917. doi: 10.1109/ICSMC.2003.1244499

Paul Robinette, Ayanna M Howard, and Alan R Wagner. 2017. Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 425–436.

Cynthia Rudin. 2018. Please Stop Explaining Black Box Models for High Stakes Decisions. *arXiv* (Nov 2018). arXiv:1811.10154v1 https://arxiv.org/abs/1811.10154

Selma Sabanovic, Marek P Michalowski, and Reid Simmons. 2006. Robots in the wild: Observing human-robot social interaction outside the lab. In *9th IEEE International Workshop on Advanced Motion Control*, 2006. IEEE, 596–601.

Fatai Sado, C Kiong Loo, Matthias Kerzel, and Stefan Wermter. 2020. Explainable goal-driven agents and robots-a comprehensive review and new framework. *arXiv preprint* arXiv:2004.09705 180 (2020).

Raymond Sheh. 2017a. Different XAI for different HRI. In *AAAI Fall Symposium-Technical Report.* 114–117.

Raymond Ka-Man Sheh. 2017b. "Why Did You Do That?" Explainable Intelligent Robots. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 628–634.

Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.

Krishna Sood. 2018. The ultimate black box: The thorny issue of programming moral standards in machines [Industry View]. IEEE Technology and Society Magazine 37, 2 (2018), 27–29.

Robert P Spunt, Meghan L Meyer, and Matthew D Lieberman. 2015. The Default Mode of Human Brain Function Primes the Intentional Stance. *Journal of cognitive neuroscience* 27, 6 (2015), 1116–1124. 9

MS Sreejith, Steffy Joy, Abhishesh Pal, Beom-Sahng Ryuh, and VR Sanal Kumar. 2015. Conceptual design of a wi-fi and GPS based robotic library using an intelligent system. *International Journal of Computer and Information Engineering* 9, 12 (2015), 2504–2508.

Paul Thagard. 1989. Explanatory coherence. *Behavioral and brain sciences* 12, 3 (1989), 435–502.

Andreas Theodorou, Robert H. Wortham, and Joanna J. Bryson. 2016. Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots. Paper presented at *AISB Workshop on Principles of Robotics*, Sheffield, UK United Kingdom. (Apr 2016).

A Narasima Venkatesh. 2019. Reimagining the future of healthcare industry through Internet of medical things (IoMT), artificial intelligence (AI), machine learning (ML), big data, mobile apps and advanced sensors. *International Journal of Engineering and Advanced Technology (IJEAT), 9, 1 (2019)*. http://dx.doi.org/10.2139/ssrn.3522960, 3014–3019.

Douglas Walton. 2006. Examination dialogue: An argumentation framework for critically questioning an expert opinion. *Journal of Pragmatic*s 38, 5 (2006), 745–777.

Douglas Walton. 2011. A dialogue system specification for explanation. *Synthese* 182, 3 (2011), 349–374.

Tong Wang. 2019. Gaining free or low-cost interpretability with interpretable partial substitute. In *International Conference on Machine Learning*. PMLR, 6505–6514.

Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. 2005. Organizing and the process of sensemaking. *Organization science* 16, 4 (2005), 409–421.

Eva Wiese, Giorgio Metta, and Agnieszka Wykowska. 2017. Robots as intentional agents: using neuroscientific methods to make robots appear more social. *Frontiers in psychology* 8 (2017), 1663.

Daniel A Wilkenfeld and Tania Lombrozo. 2015. Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education* 24, 9-10 (2015), 1059–1077.

Jeffrey C Zemla, Steven Sloman, Christos Bechlivanidis, and David A Lagnado. 2017. Evaluating everyday explanations. *Psychonomic bulletin & review* 24, 5 (2017), 1488–1500.