Research Article

Guglielmo Papagni* and Sabine Koeszegi

Understandable and trustworthy explainable robots: A sensemaking perspective

https://doi.org/10.1515/pjbr-2021-0002
received February 29, 2020; accepted June 29, 2020

Abstract: This article discusses the fundamental requirements for making explainable robots trustworthy and comprehensible for non-expert users. To this extent, we identify three main issues to solve: the approximate nature of explanations, their dependence on the interaction context and the intrinsic limitations of human understanding. The article proposes an organic solution for the design of explainable robots rooted in a sensemaking perspective. The establishment of contextual interaction boundaries, combined with the adoption of plausibility as the main criterion for the evaluation of explanations and of interactive and multi-modal explanations, forms the core of this proposal.

Keywords: explainability, interpretability, sensemaking, trust, human-robot interaction

1 Introduction

Socially assistive robots are progressively spreading to many fields of application, which include health care, education and personal services [1–3]. Whereas assistive robots must prove useful and beneficial for the users, at the same time their decisions, recommendations and decisions need to be understandable. In fact, researchers agree on the fact that social robots and other artificial social agents should display some degree of interpretability in order to be understood, trusted and, thus, used [4–6]. Concerning this connection, Miller states that "trust is lost when users cannot understand traces of observed behavior or decision" [7, p. 5]. Moreover, if the development of a trustworthy relationship is one of the main goals in social robotics, it should be considered that understanding and correctly interpreting automated decisions are at least as important as accuracy levels [4]. To this extent, "no matter how capable an autonomous system is, if human operators do not trust the system, they will not use it" [4, p. 187].

1.1 The interdisciplinary challenge of explainable robots

Automated decisions by robots already influence people's life in numerous ways. This trend is likely to be even more prominent in the future, creating a need for appropriate narratives to foster the acceptance of social robots [1–3,8,9]. Making explainable robots understandable for users with little to no technical knowledge poses a "boundary challenge" that calls for interdisciplinary efforts. In light of the growing presence of social robots and other artificial agents and of the possible consequences that their massive application might have in the future, developing an interdisciplinary approach has been deemed one of the most pressing challenges [7–11]. Following the notion of a "boundary object," the interdisciplinary issue with explainability applied to robotics can be described as "a sort of arrangement that allows different groups to work together without consensus" [12, p. 603].

For the effort to be successful, the acknowledgment and correct introduction of different disciplinary contributions are necessities that have to be met. In practical terms, this means aligning the robots' processing of information in algorithmic and "coordinate-based terms" [6], with the fuzziness of human systems. On the level of knowledge production, it requires the joint effort of several fields of research.

*** Corresponding author: Guglielmo Papagni,** Technische Universität Wien, Institut für Managementwissenschaften/E330 (TU Wien), Theresianumgasse 27, 1040, Vienna, Austria, e-mail: guglielmo.papagni@tuwien.ac.at
Sabine Koeszegi: Technische Universität Wien, Institut für Managementwissenschaften/E330 (TU Wien), Theresianumgasse 27, 1040, Vienna, Austria, e-mail: sabine.koeszegi@tuwien.ac.at

Open Access. © 2021 Guglielmo Papagni and Sabine Koeszegi, published by De Gruyter. CC BY This work is licensed under the Creative Commons Attribution 4.0 International License.

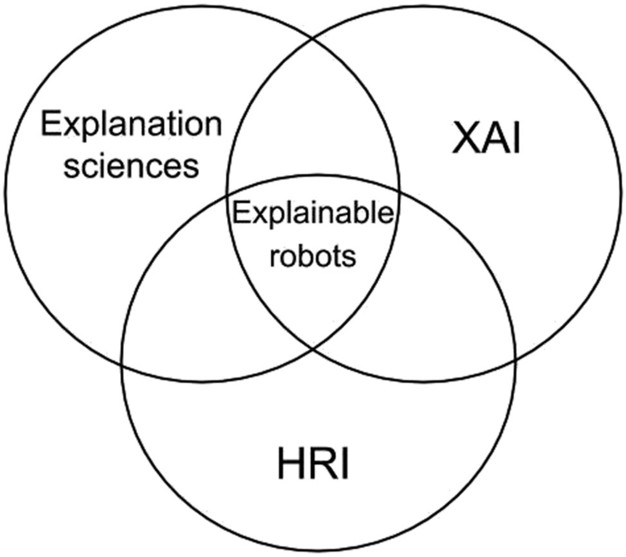

Figure 1: The interdisciplinary challenge of explainable robots.

Figure 1 shows that at the lowest level of granularity, there are at least three "disciplines" that are directly involved. Theories of explanations and causal attribution are associated with an extensive and well-studied body of literature in human interaction sciences like social psychology and philosophy [7,13–16]. Recently, these theories have received growing attention within the field of artificial intelligence, in light of the opacity of the underlying decision-making processes, particularly when these depend on machine learning models known as "black-boxes." The research area of explainable artificial intelligence (XAI), particularly in its "goal-driven" form (rather than "data-driven"), stands as the most structured attempt in making AI systems' decisions understandable also by non-expert users [8,9,17,18]. Finally, the extension of the concept of explainability to robotic technologies, especially in the forms that are meant to be used in social contexts, calls for the connection with the study of human–robot interaction (HRI) [19,20]. Each of these fields of research represents a further intersection of different disciplinary efforts. In order to successfully make social robots explainable, these dimensions need to be merged.

1.2 Making social robots explainable and trustworthy: a sensemaking approach

This conceptual article aims to advance this interdisciplinary discussion by operationalizing the core concepts of Weick's sensemaking theory [21]. Sensemaking theory is a framework from the field of organization science to describe how people make sense and understand events and, to the best of the authors' knowledge, it has not been previously applied to the domain of explainable robots.

Therefore, the article analyzes the central assumptions of sensemaking in light of the goal of maximizing non-expert users' understanding of social robots' explanations and, consequently, fostering and supporting trust development. Examples derived from the literature on different fields of application of social robots are provided to clarify and motivate the theoretical positions expressed and to show how social robots and users can benefit from the implementation of explainability in different contexts.

The article is structured into two main parts. First, in light of the connection between robots' explainability and their trustworthiness, a definition of trust suitable for the discussion on explainable robots is provided. Subsequently, the article identifies and analyzes three main issues with explainable robots. Section 2 discusses the approximate and incomplete nature of artificially generated explanations in relation to alternative forms of interpretability.

Section 3 analyzes the implicit limitations of human forms of understanding, as they also represent a challenge for the design of explainable robots. Two main issues are addressed. First, how people tend to overestimate the quality of their understanding and knowledge retention in relation to explanatory interactions is investigated. Subsequently, how access someone's intentions, beliefs and goals (upon which explanations are mostly built) is problematic is discussed, regardless of the biological or artificial nature of the agents.

Closing the first part of the article, Section 4 shows how the contextual nature of explanations is to be considered as an active force in the shaping of explanatory interactions, rather than merely a situational condition within which they occur.

The second part of the article aims to provide solutions rooted in a sensemaking perspective. Whereas the issues are approached individually, the solutions converge into a holistic model, so as to facilitate its implementation in the design phase.

Section 5 demonstrates how the contextual element can be handled via two specific features. The first refers to the role of explanations in building trust when users lack previous experience and in setting contextual boundaries for the robot's role and capabilities. Accordingly, the second relates to considering users as novices as the initial condition. The robot's mental model of the user's should only be updated after this initial phase.

As a potential preliminary solution to the approximation issue, Section 6 refers to the fact that sensemaking is driven by plausibility rather than accuracy. This relates the concepts of inference to the best explanation and explanation to the best inference. Combining these ideas, the question to be answered is "how can explanations be tailored and structured in order to maximize the chances of correct – or at least the best possible – inferences?"

Building upon this, and acknowledging the limits of understanding, Sections 7 and 8 discuss how to make sure that the best possible approximation triggers the best possible understanding. This leads first to analyzing two of the main models of iterated explanatory dialogues and, consequently, to the combination of interactive explanations with multi-modal explanations intended as "combined signals." A discussion on the conclusions and limitations of this article closes the second part.

1.3 Trusting explainable robots

There are many possible ways to conceptualize trust. Social robotics offers a unique interpretation. Andras et al. analyze different disciplinary interpretations of trust derived from psychology, philosophy, game theory, economics and management sciences [22]. Of these, they identify Luhmann's reading as one of the most comprehensive and appropriate for describing the relationship between robots and other artificial agents. Accordingly, this article defines trust as the willingness to take risks under uncertain conditions [22,23].

In principle, this conceptualization can be applied to unintentional events, where the risks to be taken are of an environmental nature, and the causes are typically natural, mechanical or societal. Conversely, when embedded in interpersonal relationships, trust exposes people to risks and vulnerability of social nature. Following this interpretation, building trust implies intentions, goals and beliefs rather than mechanical causes. People project intentionality and goal-oriented behaviors onto robots that display forms of social agency in order to try to make sense also of their actions [5]. Thus, this article refers to explanations of intentional behavior, which represents the core of interpersonal relationships, but does also apply to social robots onto which people project intentions (and goals and beliefs).

Explaining and understanding robotic decisions reduce the perceived risks involved in interacting with robots, thus fostering the development of trustworthy relationships with them. Explanations play a dual role in this context. On the one hand, they provide reasons to trust a robot when individuals lack previous experience and have not established appropriate mental models. On the other hand, they help prevent loss of trust or restore it when the robot's actions are unpredictable, unexpected or not understood [7,22].

An example can clarify the relations between the twofold role of explanations (or other forms of interpretability), trust and willingness to interact, risk and uncertainty. In an aging society, one promising field of application for social robots is elder care. One of the main goals is to help prolong elderly people's independence, supporting them in carrying out various tasks, such as medication management [24]. In such a delicate context, willingness to accept support from a robot can be hindered by uncertainties concerning the robot's reliability, particularly when the user has no previous experience. The user's uncertainty can, therefore, translate into the perception of risks, as medication management is likely to be perceived as a high-stakes domain. The perception of risks, especially during the first interaction, can be reinforced by personal predispositions to not trust novel technologies. This, in turn, could translate into fear of the robot not respecting scheduling or dosage of the medications.

Explanations about the robot's role as well as how and when it will remind the users to take their medications can therefore provide reasons to trust the robot's reliability when the interaction is initiated. Although this initial perception of risk is likely to decrease with prolonged interaction, the robot might still make unexpected recommendations, which could endanger trust if not explained. For example, some assistive robots are designed to adapt their recommendations in accordance with users' needs [25,26]. If such adaptations are not motivated (i.e., explained), users might perceive the robots as erratic and, ultimately, untrustworthy [4,6].

Understanding is not only fundamental for reinforcing users' willingness to interact with robots and other types of artificial agents. The more these social agents occupy relevant roles in our society, the more broadly they will influence our lives in general. As these types of "social robots" are being deployed in environments where many potential users have little to no understanding of how robots take decisions and make suggestions, it is necessary for them to be understood even by users without any technical knowledge.

2 Forms of interpretability: are explanations always needed?

On a general level, explanations or other forms of behavior interpretability should shed light on robots' decisions and predictions when these and the related evaluation metrics alone are not sufficient to characterize the decision-making process [7,27]. For robots' decisions to be interpretable and understandable, their inner workings must also be interpretable and understandable. In fact, the decisions represent external manifestations of the specific way robots process information. To this end, robots can be considered as embodied forms of artificial intelligence [28,29].

2.1 Direct interpretability

Practically, interpretability consists of a wide array of techniques that grant some level of access to the robot's decision-making process. In principle, not all types of artificial decision-making processes must be explained. Debugging or tracing back decisions at the level of the underlying model or even the algorithms might, in some cases, offer a sufficient degree of interpretability to grasp reasons and rationales behind a robot's actions.

These forms of interpretability are sometimes defined as "transparency," "technical transparency" or "direct interpretability" [27,31,32]. Among the models that offer this type of "readability" are shallow decision trees, rule-based systems and sparse linear models [27]. A great advantage of this type of direct inspectability is its higher transparency, which increases the possibility of detecting biases within the decision-making process and implies lower chances of adversarial manipulation. This, in turn, has a positive impact in terms of fairness and accountability [32].

Sun reports on an experiment seeking to classify elderly users' emotional expressions using tactile sensors installed on a robotic assistant so that it can give an appropriate response [33]. Two of the classifiers used to identify the participants' emotional expressions are a temporal decision tree and a naive Bayesian classifier. Even though some of these models can be accessed directly, this form of accessibility to the decision-making process requires some technical expertise and is likely to be mostly useful for expert practitioners and developers [34].

A problem emerges as one of the key criteria of this article is to make robots' decisions understandable for all types of users, including non-expert users. In terms of everyday interactions with social robots, this type of users is likely to represent the majority [4,35]. In this case, it should be assumed that they have little-to-no knowledge of how even relatively simple and intuitive models work.

The use of robotic companions like the one in the aforementioned example, capable of recognizing emotions among other tasks, can be expected to increase markedly in the future. Hence, situations might arise in which the response given by the robot does not match the emotion expressed by the user. The latter might want to know why the robot responded inappropriately to an emotional expression. Whereas an expert practitioner can benefit from direct forms of interpretability, the same cannot be automatically said about non-expert users. If anything can be assumed at all, it is that for non-expert users, this type of accessibility would be too much information to handle (i.e., "infobesity" [35] or require too much time to be understood [36]). Considering the "limited capacity of human cognition," there is a chance that providing this type of information would result in cognitive overload [27, p. 35].

Returning to the example, in the best case, the user would simply fail to understand why the robot provided the wrong emotional response. Alternatively, failing to understand the robot's action might cause unsettling and erratic feelings which, in turn, could lead to a loss of trust in the robot [4,6]. Moreover, direct forms of accessibility to the decision-making process are not available for all types of models implemented in social robots.

2.2 "*Post hoc*" interpretability

Explanations generated "*post hoc*" represent an alternative type of interpretability. Since seeking and providing explanations is a fundamental form of "everyday" communicative social interaction, this solution seems to be more useful for non-expert users [7,37].

Popular complex models like deep neural networks (often labeled "black-boxes") process inputs to produce outputs in opaque ways, even for expert users. Therefore, in order for their predictions, decisions and recommendations to be understandable, a second simpler model is usually needed to clarify, through text-based explanations or other means, how inputs are processed into outputs [38]. This form of interpretability can sometimes also be applied to models that are typically considered to be directly interpretable [27,30].

The information is mostly generated in human-friendly terms, and this is the fundamental reason that makes explanations more suitable for the needs of non-expert users. For a communication act to be defined as social, the information conveyed by the robot must be socially acceptable, rather than too technical [31,37]. Therefore, in the considered case of emotion recognition, if the robot were to misread a user's emotional expressions, the user would likely expect a justification conveyed in a socially acceptable and understandable form. For example, the robot might explain that it has mistakenly identified and classified certain parameters of the user's emotional expression in a text-based form. As the article will discuss further on, other channels of communication can also help provide socially acceptable and tailored explanations.

2.2.1 Explanations as approximations

A problem that rises with explainability is that the second model (known as the explainer) provides insights into how the complex model works, but the result is merely an approximation of the original decision-making process, rather than a truthful representation of it [32,38]. Thus, the resulting explanations have a varying degree of fidelity to the actual decision-making process depending on factors, such as the type of task, the models implemented in the robot and the type and depth of the explanation.

Wang highlights the twofold essence of the problem:

> First, explainers only approximate but do not characterize exactly the decision-making process of a black-box model, yielding an imperfect explanation fidelity. Second, there exists ambiguity and inconsistency in the explanation since there could be different explanations for the same prediction generated by different explainers, or by the same explainer with different parameters. Both issues result from the fact that the explainers only approximate in a post hoc way but are not the decision-making process themselves. [38, p. 1]

Post hoc interpretations are therefore problematic for several reasons. Since explanations are open to interpretation, they can be simply misinterpreted by the explainee. More dangerously, they can hide implicit human biases in the training data or even adversarial manipulations and contamination [32]. Explanations might therefore be coherent with the premises and with the data used to generate them; yet, those premises are wrong [39].

Nevertheless, despite the fact that such explanations do not precisely convey how the robot's underlying model works, they still appear to be the most suitable option for social robots. Despite their approximate nature, most of the times explanations still convey useful information, which can be tailored in user-friendly terms. To this extent, *post hoc* interpretability is the strategy that people also use to make their decisions interpretable to others [27]. Moreover, if social robots are in principle designed to be understandable by users with no technical knowledge, experienced users and developers will also be able to make sense of these explanations. Further access to a deeper level of information processing can still be granted if the user requests it, depending on the availability of the implemented models [30].

In conclusion, if it is true that non-expert users can benefit from robots' explainability, then direct forms of interpretability pose a problem when it comes to the fairness and accountability of robots. Accordingly, the question to be answered is how to ensure the best level of approximation possible. This implies explanations that are coherent with the actual decision-making process, understandable and meaningful for the user and, perhaps more importantly, disputable.

3 Limits of understanding

Successful explanations are the result of contextual joint efforts to transfer knowledge and exchange beliefs [7,36,39]. For the explainer, this implies crafting explanations that are potentially good approximations of the actual decision-making processes, while on the other side of the information transfer, the explainee's knowledge must be successfully updated. Thus, Section 3.1 identifies and discusses two main cognitive elements that hinder successful understanding.

3.1 The problem of introspection

People tend to overestimate the amount and quality of the retained information. Keil states that the first introspection is not very reliable when it comes to "explanatory forms of understanding." More generally, people's understanding of how things work, especially at a naive level, is far less detailed than it is usually thought [40].

In social psychology, this phenomenon is connected to the concept of the “introspective illusion” [41]. Consequently, when it comes to explanations’ reception, even when the explainee consciously declares that they have reached a sufficient level of understanding, this is not always the case. Keil terms this the “illusion of explanatory depth” [40]. It might also be that, despite being aware of not having reached a sufficient level of understanding, the explainee still claims the opposite. This more conscious appraisal of knowledge retention would likely be due to other reasons, such as the desire to meet someone’s (e.g., the explainer’s) expectations. In both cases, the result is that, when people are questioned about their understanding of something that has been explained to them, an incorrect estimation of retention quality emerges.

If knowledge retention is not tested, such possible misinterpretations can remain unacknowledged. In sensemaking theory, this issue is expressed with the notion that people tend to take sensemaking for granted, whereas this is a subtle, ongoing process that should be lifted from an implicit and silent to an explicit and public level [21].

3.2 Inaccessibility of other’s intentions (and minds)

The phenomenon of overestimating one’s own understanding also plays a role in creating wrong mental pictures of others, in the sense of a folk-psychology theory of mind [42]. In other words, it influences people’s ability to predict others’ behavior and the reasons, intentions and goals behind it [40]. In accordance with the enduring philosophical issue known as the “problem of other minds” [43,44], the question arises as to how we can be sure that we have understood other people’s intentions and beliefs upon which explanations are generated.

Considering this issue in light of the “information asymmetry” proposed by Malle, Knobe and Nelson, it becomes clear that this issue can occur also in the field of social robotics [45,46]. The authors note that an observer (that in the case of an explanatory interaction would be the explainee) would not provide the same explanation for an action as the actor who performed it. Generally, the difference in explanations is because the observer has to infer the other actor’s intentions from their behavior, precisely because these intentions cannot be accessed directly [7, p. 19].

People tend to refer to robots as social actors despite their artificial nature, at least partially because they perceive intentionality behind robots’ actions [5,7]. This implies that when people interact with robots perceived as having reasons, intentions and goals behind their actions, the concept of “information asymmetry” overlaps with the inaccessibility of the robot’s intentions. Therefore, when a robot makes a suggestion, the recipient has to initially infer what might be the reasons for this recommendation. From the perspective of explainable social robots, the risks of users failing to introspectively assess their retention of knowledge and to infer the robot’s intentions should be considered default conditions that can never be ruled out [47].

As previously mentioned, a growing field of application for social robots is elderly care. Among other tasks, assistive robots are meant to help fight loneliness and prolong elderly people’s independence by performing various daily tasks such as monitoring health and medication management, or supporting with household duties [1,24,48]. It has been reported that elderly users might greatly benefit from the company of these types of robots, but only if they prove to be efficient and useful [1]. For instance, already now IBM’s Multi-purpose Eldercare Robot Assistant (IBM MERA) is designed to learn users’ patterns and habits and adapt its care suggestions accordingly through a combination of environmental data gathered in real time through sensors (e.g., located on the floor, walls and ceilings) and cognitive computing [49,50]. Similarly, other assistive robots include functions for detecting obstacles and clutters on the floor in 3D as well as other adaptive behaviors [25,26].

As assistive robotic technologies become more sophisticated and adaptable to changing environments, their recommendations will become even more nuanced. Since successful explanations imply successful transfer of knowledge [7,39], ensuring that users can understand and make sense of these adaptable care services and infer the right intentions behind robots’ actions is a high priority in order to avoid potential negative consequences. Moreover, a potential positive consequence of correctly inferring reasons and understanding explanations is that it allows the user to check whether the explanations are based on flawed or accurate reasons. Such situations require making the successful (or unsuccessful) understanding of an explanation explicit, as the sensemaking framework proposes [21].

4 Context dependence

The previous sections have discussed how successful explanatory interactions pose challenges for both the explainer and the explainee. Typically, the structure of explanations includes these two elements and the message to be conveyed (i.e., the "explanandum"). However, a fourth element must be considered in light of the fact that explanatory interactions do not occur in neutral environments. Rather, they are contextualized events, and in the case of explainable social robots, the context in which explanations are sought out and provided is predominantly social. As Weick, Sutcliff and Obstfeld note, to understand how people make sense of events, the focus needs to be shifted "away from individual decision makers toward a point somewhere 'out there' where context and individual action overlap" [21, p. 410].

Malle identifies two main reasons why people seek explanations for everyday behaviors: to find meanings and to manage social interactions [42]. This twofold approach is rooted in the folk theory of mind and behavioral framework in which explanations are embedded. In the sensemaking framework, finding meanings in a context means bringing order to the chaotic stream of both intentional behaviors and unintentional events that constitute the social environment [21]. People do this through the ascription of reasons and causes. Before explaining intentional behaviors, these need to be discerned from unintentional events, which typically have more mechanistic explanations (e.g., natural phenomena [7]).

At the same time, as sensemaking is a social and systemic event, influenced by a variety of contextual and social factors, a contextual analysis can help to identify the conditions that made an action possible [21]. Therefore, the context cannot be reduced to the traditional dichotomous relationship between a person and a situation (attribution theory). Explanations as contextual events involve finding meanings in a co-constructive process, where the context is an overarching and active force that influences how the interaction takes place (rather than being only "situational"), as shown in Figure 2. It is within this context that explanations function as a tool people use to manage communication, influence, impressions and persuasion with each other [32,51].

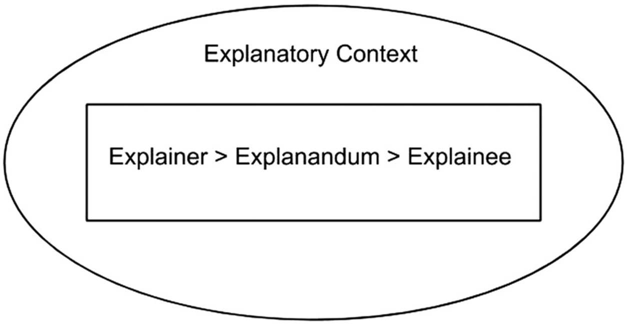

Figure 2: Explanations as contextual events.

4.1 Different contexts imply different explanations

One contextual element that can deeply influence how robots can and should explain their actions is time availability. This refers to how much time the user is able or willing to invest in receiving (i.e., listening to and understanding) an explanation. The impact of this variable varies widely depending on the field of application of social robots.

For instance, several studies investigate how robots can be used to assist with carrying out tasks in libraries [52–54]. Some of these robots can recommend potentially interesting reading material to users based on feedback and reviews from other users [54]. In such a scenario, a library customer might want to know whether a recommended book or periodical is worth reading and, therefore, decide to spend some time figuring out whether she would like the book or magazine before starting to read it. Hence, the user would benefit from a relatively detailed explanation. On the contrary, in other situations where decisions and actions must be taken quickly, externally imposed time constraints can force the explainable robot to combine speed with a sufficient level of detail.

One such case concerns robots involved in rescue missions, as discussed in [28,29,55]. According to Doshi-Velez and Kim [36], explanations are not required when no notable (typically negative) consequences are at stake. Perceived potential consequences represent another contextual element that can deeply influence explanatory interactions. In the case of the robotic librarian, a potentially negative consequence the user might identify is that she decides to read a book that he or she does not like. Although the user's decision on whether or not to read the book is a low-stakes one, she might still want to invest some time to query the suggestion further before deciding. Conversely, situations where robots are involved in rescue missions or military operations pose a particularly difficult challenge. In such case, the time the user can invest is likely to be low, while the potential consequences of a wrong choice might be highly negative. As Section 6 discusses, applying plausibility over accuracy as a key criterion for explainable robots provides a potential solution to this problem.

In conclusion, this section sought to demonstrate that contextual elements should not only be considered as environmental conditions within which explanatory interactions take place. Because the context can directly or indirectly influence how explanations are conveyed and received, the active role it plays must also be considered at the implementation level as the other issues identified in the previous sections.

Thus, the following sections address the issues identified to operationalize sensemaking-based intuitions aimed at maximizing non-expert users' understanding of robots' explanations. The analysis follows three main directions. Section 5 shows how treating "users as novices" and explicitly constructing contextual boundaries can support users in dealing with the contextual nature of explanatory interactions. In accordance with this, Section 6 emphasizes on the (agreed upon) plausibility of explanations as the evaluation key criterion. Sections 7 and 8 investigate "interactive" and "multimodal" explanations, respectively, as related strategies for dealing with the approximate nature of explanations and the limitations of human understanding. Although their potential positive impact has been recognized, to the best of the authors' knowledge, they have never before been combined in an organic model [9,56].

5 Users as novices and contextual boundaries

When providing explanations, people try to tailor them to the person asking them [57]. In other words, explanations are adapted to the (possessed) explainee's mental model, especially to the perceived level of expertise [57]. As discussed in Section 4, the context within which explanations are requested and provided plays a fundamental role in shaping the interaction. These elements influence several parameters of the explanation structure, including the level of detail/depth, the material included in the explanation and the communication strategy, including the level of "technicality" that can be used [57].

Both models for interactive explanatory interactions discussed in Section 7 assume that initially the parties involved have some degree of shared knowledge about the topic of the explanation [39,58,59]. Accordingly, the explanatory interaction begins when one of them (i.e., the explainee) detects an anomaly in the other's account and, thus, requests an explanation [39,58,59]. In everyday interactions, the parties involved are likely to share at least some common knowledge about the events being discussed (and explained). However, this can be problematic when it comes to explainable robots.

5.1 Explainable robots in the wild

In many experimental cases in elderly care, when the robot is introduced to the users, it is made clear that they can count on its support in carrying out tasks. As social robots are also meant to operate "in the wild," many situations will represent a "first time" and "one-shot" interaction. In such scenarios, it is important that contextual boundaries are set and proper mental models established so that interactions can proceed smoothly. As discussed in Section 4, contextual limitations can influence the development of an explanatory interaction with robots. Consequently, the need to co-construct the context should be taken into consideration. Thus, Cawsey suggests that, at the beginning of the explanation, explainees should be treated as novices, and mental models adapted accordingly as the interaction progresses [57].

In case of "first time" interactions in non-controlled environments, the co-construction of the context with a social robot already implies explanations for why the robot has made an approach and is willing to interact. Following Miller's argument, people's requests for explanations mainly occur in the form of why questions [7]. During an initial interaction, these questions might be implicit.

For instance, such a situation might occur with robotic shopping mall assistants, as discussed in [30,60,61]. When the robot approaches a potential customer, she might wonder why the robot is talking to her. In this situation, if the robot were to opt for proactive behavior, the establishment of context boundaries would correspond to an explanation of what the robot's role is and why is it approaching this particular potential customer. By introducing itself and proactively clarifying its role, the robot is answering the user's potential and typically implicit question of why the robot wants to interact. In turn, following the "foil argument," by explaining that its role is to provide shopping recommendations, help navigating or other similar tasks, the robot automatically rules out other possibilities [7]. If the potential customer would not have asked this question explicitly, the robot's explanation lifts doubts and knowledge from an implicit and private level to an explicit and public level [21].

Setting the contextual boundaries as described could make the robot aware of other contextual elements, such as whether the customer has time to invest in considering to the robot's shopping recommendations. More importantly, in line with the dual relationship between trust and explanations described in the first section, this approach to contextualization is appropriate for situations in which inductive trust has not been established yet because previous experience is lacking [22].

As the robot approaches the potential customer, perceived "information asymmetry" would likely be at its peak, as the user might not be able to infer the robot's intentions [45]. By explaining its role, the robot can minimize this phenomenon and provide the user with reasons to inform their decision of whether or not to interact with the robot and trust its shopping recommendations. At this point, the user can express her understanding and intention to either continue the interaction or not. The robot is therefore able to update its mental model of the user accordingly without necessarily needing to ask further questions, as suggested in [57]. Finally, referring to the idea that each interaction triggers a new sensemaking request [21], during further approaches to the same potential customer, the robot should be able to investigate, perhaps by questioning the user, whether its previous mental model is still valid.

5.1.1 Non-verbal cues

In the considered scenario with the shopping assistant, another element can support setting initial contextual boundaries. Specifically, the robot can let the potential customer know that it is approaching her through non-verbal cues (e.g., body posture and movements, gaze, graphic interfaces and light signaling).

At the entrance to a shopping mall, the interaction context can be crowded. The user might not understand immediately that he or she is the target of the robot's attentions. In such cases, it has been demonstrated that non-verbal behavioral cues and signals can foster the perceived social presence of the robot and the users' engagement and, therefore, support the establishment of contextual boundaries before the verbal interaction begins [62–65]. Accordingly, it has been demonstrated that such complementary channels can help users make sense of robots' intentions and therefore support the initial generation of a correct mental model of the robot. Eventually, the user might realize before any verbal interaction that the robot is a shopping assistant and, thus, immediately decide whether to avoid or proceed with the interaction.

6 Plausibility over accuracy

Section 2 demonstrated that, compared to other types of interpretability, making robots explainable has better chances of maximizing non-expert users' understanding of robotic decisions and suggestions. *Post hoc* interpretability is also the strategy that people use to shed light on their "biological black boxes," although the process is not always successful. Nevertheless, in most cases people manage to convey meanings and information in explanatory interactions.

How can explainable robotics make use of this to develop a strategy for handling the approximate nature of explanations? A possible solution is rooted in a fundamental element of the sensemaking theory. As a process, "sensemaking is driven by plausibility rather than accuracy" [21, p. 415]. This is in line with the pioneering work of Peirce on abductive reasoning [66] in the field of explanation science. As a cognitive process, explaining something is better described in terms of abductive reasoning, rather than inductive or deductive [7]. Like inductive reasoning, the abductive reasoning process also involves proceeding from effects to causes. However, in deriving hypotheses to explain events, abductive reasoning assumes that something "might be," rather than just "actually is" [7,66].

Applied to the inference process for explanations, this intuition has been translated as "inference to the best explanation" [67]. Whereas in this case, the emphasis is on explanations as a product of the inference process, Wilkenfeld and Lombrozo interpret the processual act of explaining as "explaining for the best inference" [68]. This leads them to posit that even when a correct explanation cannot be achieved, one's cognitive understanding of the process can still benefit [68]. Beyond the possible "cognitive benefits" of even inaccurate explanations, what is more important for this article is the notion that people do not seek to obtain "the true story." They rather seek out plausible ones that can help them grasp the possible causes of an event [21].

Abductive reasoning offers a reading key where plausibility emerges as a key criterion for the selection of a subset of causes that could explain an event. In light of this, the goal for explainable robots shifts from providing "objectively good" to "understood and accepted" explanations. In other words, explanations should trigger "the best inference" possible about the causes of robots' decisions. Recalling the idea of the co-constitution of meanings in sensemaking theory, the plausibility of an explanation should also be considered a joint (contextual) achievement.

6.1 Explanatory qualities

What properties an explanation should have is debated within the field of explanation science. In fact, "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" [13, p. 27].

Some researchers identify simplicity as a desirable virtue. If explaining a phenomenon requires fewer causes, it is easier to grasp and process the explanation [69,70]. Other researchers argue that completeness and complexity enhance the perceived quality and articulation of an explanation [69,71]. An explanation of internal coherence and coherence with prior beliefs is generally considered a further relevant quality [7,14,70,72].

Since explanations are contextually co-constructed events, the joint achievement of plausibility seems to overrule the question of simplicity and complexity. According to the context within which explanations are requested, their joint evaluation as plausible includes whether the amount and complexity of information provided were satisfying in selecting a subset of causes but not overwhelming or too elaborated. Such a solution might help in dealing with potentially hazardous situations, as described in Section 4. Moreover, for an explanation to be agreed upon as plausible, it must be coherent with prior beliefs (particularly of the explainee), or, at least, potential contrasts between the new pieces of information and prior beliefs must be resolved.

Although using plausibility as a key criterion for how explainable robots should structure their explanations seems theoretically valid, a problem arises in cases when an explanation is plausible, but nevertheless based on incorrect premises [39,73]. For instance, if an assistive robot were to suggest that a user avoids a certain path through the house and motivates the suggestion by explaining that it detected an obstacle, then this reason might be considered plausible. However, if the premise is wrong (e.g., the obstacle detected is a new carpet), the plausibility is rooted in an inaccurate foundation. Importantly, this principle must be implemented together with a strategy for challenging the explanation in case it sounds anomalous. In Sections 7 and 8, possible solutions are proposed to ensure the best level of approximation and to maximize users' understanding.

7 Interactive and iterative explanations

Interactive explanations have already been successfully developed into models and tested [39,57–59]. This section discusses two of the most elaborate recent models for explanatory interactions with artificial agents in light of the issues identified in the previous sections and with respect to their application with explainable robots. The first is Walton's system for explanations and examination dialogues [39]. The second is the grounded interaction protocol for XAI recently proposed by Madumal et al. [58,59].

These two approaches are considered because both take into account the end users' perspective as a central feature; however, they do so in different ways. The former takes a more theoretical approach, while the latter is based on actual data collected from human-human and human-agent interactions. Nevertheless, while these two approaches can be interpreted as complementary, they both lack certain elements that are central for users' sensemaking. As discussed in Sections 4 and 5, one of these missing elements is a consideration of the contextual nature of explanations (particularly, in first interactions).

7.1 Context consideration in interactive explanations

Before analyzing the relevant features of the models, two reasons are identified to support the establishment of the initial contextual condition as a means of building a solid foundation for further interaction. First, if the context is not established initially, possible misunderstandings can emerge as the interaction progresses.

At this point, it becomes more difficult to trace back what was not understood. Potentially, this initial setting can prevent or help to more quickly identify the causes of what Walton named the problem of the "failure cycle," which occurs when the examination dialogue cannot be closed successfully (i.e., when the explainee repeatedly fails to understand) [39, p. 362].

Furthermore, it has been argued that lifting knowledge from the private and implicit level to the public, explicit and thus usable level is a fundamental element of shared sensemaking and should not be implicitly assumed [21]. Walton seems to acknowledge this. He notes, "to grasp the anomaly, you have to be aware of the common knowledge" [39, p. 365]. Moreover, again, when describing deep explanations as the most fitting for the dialogue model, he states that "the system has to know what the user knows, to fill in the gaps" [39, p. 365]. Nevertheless, he makes clear in the development of his model that the system makes assumptions about the user's knowledge.

After the initial establishment of contextual boundaries and conditions of explanatory interactions, this part of the article mainly considers explanations as embedded in prolonged interactions. For instance, this would be the case with assistive robots used for elderly care, particularly, as they are also meant to become companions to fight against loneliness and therefore object of long-term interactions [74].

7.2 Anomaly detection

Once the interaction context and initial mental models have been mutually established, it might happen that a user requests an explanation from a social robot, typically for unexpected or unpredictable behaviors.

In such cases, the explanatory interaction is usually triggered by an "anomaly detection," as termed by Walton [39]. Similarly, in the work by Madumal, Miller, Sonenberg and Vetere, the identification of a "knowledge discrepancy" is the initial condition for an explanatory dialogue [58,59]. This step reflects the second approach to the relationship between explanations and trust analyzed in the first section. There, it was described how unexpected or unpredictable robotic behaviors, if not explained and understood, could undermine trust in the relationship.

For instance, this can be the case with advanced assistive robots like IBM's MERA, which is capable of monitoring the user's pulse and breathing [1,49]. If the robot were to detect variations in these parameters, it might recommend that the user take a rest. Different elements can trigger the detection of an anomaly. This is also linked to the perception of "information asymmetry." What changes is the robot aware of (that motivate the suggestion) but the user is not? Perhaps the user has not yet consciously recognized the variations identified by the robot, or perhaps the robot usually recommends that the user take a rest at different scheduled times throughout the day. In any case, the user might find the suggestion anomalous and this would likely trigger an explanation request.

Referring to the discussion in Section 5 about proactive robotic behavior in the establishment of an explanation context, the robot does not necessarily have to wait for an explicit request. With simple, introductory *a priori* explanations, the robot can act in advance of the suggestion, hence reducing the chances of an anomaly detection: "I have detected that your heart and breath rates are above the norm. Maybe you should take a rest." Generally, such proactive explanations can be presented in compliance with rules of conversation like the four "Gricean maxims" [75] and can be useful in reducing the need for questions from the explainee, although this is not a guarantee against further discussion.

7.3 Explanations and argumentation

The model by [58,59] includes the option of embedded argumentative dialogues, which might deviate from the original question and are treated as cyclical. Although his model draws upon argumentation theories, Walton classifies the case of a further overlapping dialogue (meaning one that does not contribute to the original one) as an illicit dialectical shift. Following Weick, Sutcliffe and Obstfeld, sensemaking is best understood at the intersection of action, speech and interpretation, which means that, in real-life scenarios, argumentation often occurs within the same explanatory dialogue as a way to progressively refine understanding [21].

Considering the previous example of IBM's MERA, while the robot is explaining its suggestion, the user might still ask further unrelated questions, for example, whether the irregular parameters detected match the symptoms of a stroke. In such cases, the robot should be able to address new questions without necessarily considering the previous ones as closed. For this reason, as shown in Figure 3, the option of internal and external loops coding argumentation dialogues that are related and unrelated to the original question, respectively, seems to offer a more realistic perspective than merely labeling a shift as illicit [39].

7.4 From explanation to examination

Of particular interest for comparing the two models is the choice of either returning or not to an "examination phase" within explanatory dialogues. Madumal et al. criticize the choice to implement an examination phase described in [39]. They define this resorting to embedded examination dialogues in Walton's model as "idealized" because it is not grounded on empirical data. Therefore, according to the authors, it fails to capture the "subtleties that would be required to build a natural dialogue for human-agent explanation" [59, p. 1039]. Accordingly, one of the main reasons for the alternative approach adopted by Madumal et al. [59] is that, in most cases of everyday explanations, there is no explicit test of the explainee's understanding.

As the authors note, their focus is on creating a "natural sequence" in the explanation dialogue. Therefore, the examination phase is fundamentally replaced by "the explainee affirming that they have understood the explanation" [59, p. 1038]. In light of the issues discussed in Section 3, the explainee's affirmation should not in principle be considered as a sufficient criterion, and the possibility of overestimating one's understanding and knowledge retention should be explicitly addressed. Therefore, the main limitation of the model by Madumal et al. [59] is that it lacks evaluation strategies to assess whether the explanatory interactions are actually successful.

7.4.1 Examination of robotic explanations

A second reason supporting the implementation of examination dialogues is that this phase does not only test the explainee's understanding. Perhaps, more importantly, this type of dialectical shift also provides a tool to investigate the quality of the explanation itself. In other words, the aim of examination dialogues is generally to gather insights into a person's position on a topic in order to either test understanding or expose potential inconsistencies. Hence, the target of the examination can also be the explainer's account, and weak points of an explanation can be identified in the form of a request for a justification for the claims made [39,73,76]. So interpreted, examination dialogues represent a useful tool for cases in which explanations sound plausible but are grounded in inaccurate premises or information.

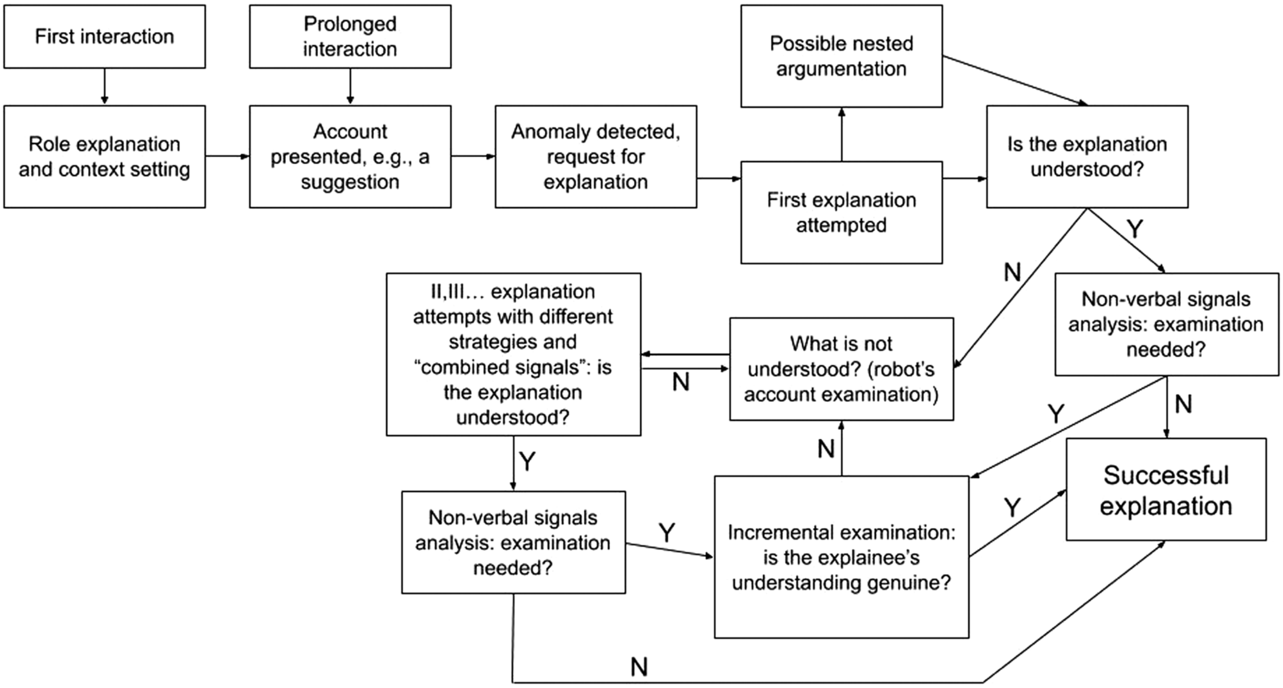
First interaction
Prolonged interaction
Possible nested argumentation
Role explanation and context setting
Account presented, e.g., a suggestion
Anomaly detected, request for explanation
First explanation attempted
Is the explanation understood?
Y
N
II,III… explanation attempts with different strategies and "combined signals": is the explanation understood?
N
What is not understood? (robot's account examination)
Non-verbal signals analysis: examination needed?
Y
N
Y
N
Successful explanation
Non-verbal signals analysis: examination needed?
Y
Incremental examination: is the explainee's understanding genuine?
Y
N

Figure 3: Explanatory dialogue model.

7.4.2 Issues of interactive explanations

Whereas implementing a shift from explanation to examination dialogues appears to be a valid solution to maximize users' understanding and the veracity of explanations, two further possible issues emerge. The first issue, as Walton reports, is directly related to the implementation of examination dialogues, which can sometimes become quite aggressive [39, p. 359].

The goal of examination dialogues is to test understanding, not to interrogate the explainee or make her feel uncomfortable or overwhelmed. If the explainee perceives the robot as hostile, the robot's trustworthiness and further interactions might be compromised. The second issue is what Walton labels the problem of "the failure cycle," which can in principle affect any type of explanation, regardless of whether or not there is a switch between explanation and examination.

One possible way for robots to proactively deal with the first issue is through social signal processing and the analysis of non-verbal cues. In social interactions, people use a wide variety of alternative channels to express themselves beyond verbal communication. There are at least two complementary reasons for social robots to use and process non-verbal communication and social signals in order to stimulate and support explainees' understanding.

As discussed in [62], non-verbal behaviors provide fundamental support to achieve "optimal" interactions in terms of engagement. The more flexible and inclusive the robot's modalities of communication, the easier it is for people to correctly read and follow robotic behaviors [62]. The understandability and persuasiveness of a robot's explanation can therefore be improved by displaying such cues.

Perhaps more important from the perspective of ensuring understanding is that the robots can analyze the same types of non-verbal behaviors expressed by the users. For instance, in order to decide whether to examine the explainee's knowledge retention, a robot could ask whether an explanation was understood and analyze the non-verbal signals accordingly. Pérez-Rosas et al. report how people tend to show specific signals when they lie and how these elements can be captured by computational methods [77]. Parameters like gaze direction and facial expressions, posture, gestures and vocal tones can be analyzed by the robot to determine whether or not the explainee's claims are genuine. Such a strategy would likely prove more efficient when users consciously claim to have achieved a deeper understanding than they actually have.

If, instead, the explainee genuinely, but erroneously, believes that they have understood an explanation, their non-verbal signals would be more nuanced. Furthermore, certain non-verbal signals (such as gaze movements [78]) are not always reliable and should not be taken by the robot as absolute evidence, but rather as useful clues as to whether an examination might be necessary to lift possible implicit misunderstandings to an explicit level [21].

7.4.3 Questioning the explainee

Even if the robot determines that an examination dialogue is needed to test the user's understanding, the questions should not be perceived as overwhelming or hostile. Walton proposes a "Scriven's test" [39, p. 357] in the form of a dialectical shift in which questions are posed to the explainee. Although these probing questions should be related to the topic, they can also help highlight connections that were not explicit in the explanation dialogue. Furthermore, as noted above, the dialectical shift should also allow the explainee to analyze the explainer's account, determine whether the explanation is sound and plausible or whether there are weak points that might uncover inaccurate information.

Walton's model does not specify how this questioning phase should be structured (e.g., how many questions should be asked). Specifically, it might be problematic for the user to have to answer many questions in terms of perceived hostility, particularly for very low-level explanations.

With reference to the fact that explanatory interactions are embedded in and influenced by specific contexts, a possible solution is to proceed incrementally, following the progression of the explanation. In other words, if the explainee expresses her intention to obtain deeper and more detailed insights on the reasons and intentions behind an explanation, the robot can assume that she is willing or able to invest time in understanding the explanation. Alternatively, as analyzed in Section 4, there might be practical reasons why the interaction cannot go on for too long. Following the sensemaking idea of focusing on the contextual conditions that make the interaction possible [21], examination dialogues and explanatory interactions more generally should be calibrated to these specific contextual conditions, rather than being decided in advance.

As social robots become more sophisticated and connected, such a functionality will likely become easy to implement. Currently, similar capabilities can sometimes be achieved through systems for "questions and answers" dialogues between the robot and the user. For example, IBM's MERA offers an interface to interact with IBM's Watson Dialogue Q&A, through its current embodiment (in the form of a SoftBank Pepper robot) and cloud connections [49].

8 Multimodal explanations and the problem of the "failure cycle"

The last issue to be dealt with to maximize users' understanding of robotic explanations is what Walton identifies as the "failure cycle" [39]. In practice, this translates into the explainee repeatedly failing to understand an explanation. Whereas the author acknowledges that in some cases external limitations and intrinsic constraints can affect the number of times that an explanation can be reiterated, he suggests rephrasing the explanatory message as a possible solution before moving on to the explanation closing stage. Nevertheless, he does not explicitly mention how explanations should be rephrased [39]. This section proposes two possible complementary solutions.

Typically, social everyday explanations take the form of natural language acts of communication. As such, according to Hilton, they should follow the rules of co-operative conversation [37]. Specifically, the author refers to "Grice's (four) maxims of conversation," which are considered a useful and "implementable" model for explainable robots and other artificial agents [7]. These are quality, quantity, relation and manner [75]. The first refers to saying only things that are believed to be true with sufficient certainty. The second can be interpreted as trying to avoid an overwhelming amount of information, i.e., seeking the right quantity. The third refers to what Hilton identifies as a good social explanation, i.e., it must be relevant to the context. Finally, the fourth refers to the mode of presenting information, in order to be clear (avoiding obscurity and ambiguity), brief and orderly [7,75,79].

Several of these qualities have already been addressed in this article. As the failure cycle mostly refers to explanations that are not understood despite the robot's clarification attempts, the first possible solution proposed here refers directly to the fourth maxim.

8.1 Alternative verbal strategies

One strategy that can be adopted to "rephrase" an explanation is to amplify the range in terms of depth and type, as suggested by Sheh [30]. The author analyzes the possible combinations of 3D levels of depth with five typologies of explanations [30]. The relevance of Sheh's approach mainly lies in the fact that he adopts an HRI perspective to analyze the options offered by machine learning models. This implies that the different types of explanations and the depth level that can be displayed are sorted according to the models implemented in the robot.

For example, Sheh analyzes [30] the case of a robotic shopping mall assistant that is asked questions about product recommendations. The explanations provided by this type of social robot, he notes, "are mostly for the purpose of satisfying the user's curiosity and as a way for the agent to further engage in dialog with the user. Post-Hoc explanations may be quite acceptable at Attribute Only or Attribute Use levels" [30, p. 117]. Referring to the potential need to rephrase an explanation, the "Attribute Only" or "Attribute Use" levels of explanation represent different potential strategies. In the former case, the explanation reveals whether the robot's decision is based on considering reasonable factors, rather than on irrelevant factors. In contrast, explanations at the "Attribute Use" level "include the implications of the values of their attributes" [30, p. 116].

8.2 Combined signals

Complementary to presenting different typologies of explanations and at different levels, multi-modality or "combined signals" [80] represents a second promising yet underrepresented direction. Anjomshoae, Najjar, Calvaresi and Främling derive six modalities of providing an explanation from the analysis of 108 core papers [56]. Text-based natural language explanations cover a significant part of the spectrum. The other explanation modalities are, in order of importance, visualization, logs, expressive motions, expressive lights and speech [56].

This does not mean that, in order to be understood, a robot should display all available information in all available formats at once. In fact, if the alternative communication strategies would be displayed all at once, their messages would overlap, likely resulting in cognitive overload. Rather, it means that while different typologies can be integrated in a complementary and supportive way (as “combined signals” [80]) within the same robotic explanation, the decision should still be in the user’s hands.

For instance, referring to the possibility of a failure cycle, the user might request a more detailed explanation that also includes graphic material. This option is described in [81], where graphic explanations for the recognition of images are accompanied by text captions describing fundamental parameters influencing the recognition process. The results indicate that such an explanation format enhances the likelihood of users grasping the reasons behind predictions.

Similarly, an assistive robot might use combined signals to improve the quality of its explanations. For example, if an elderly assistant robot recommends a user to take rest after detecting increased heart and breathing rates, it could corroborate the effectiveness of the message by displaying a graphical comparison between normal and unusual rates.

In other cases, single channel explanations (as opposed to text-based explanations) can even be a better choice overall. For example, in their work on robotic behaviors, Theodoru, Wortham and Bryson claim that, since artificial agents can take a great number of decisions per second, providing information verbally might be difficult for users to handle. In the case of reactive planning considered by the authors, they suggest that a graphical representation is more efficient and direct for making the information available even for less-technical users, while preventing them from becoming overwhelmed [35].

In conclusion, one might argue that it is impossible to ensure success in each and every explanatory interaction, as an explainee might still fail to understand the information conveyed through an explanation. Just like in human interactions, issues like the failure cycle might not be completely solvable. Nevertheless, when it comes to robots, there is a chance to address these problems ahead of practical implementations.

9 Conclusions and limitations

As social robots are becoming a daily reality, it is important for them to be able to explain their decisions in user-friendly terms. Therefore, this article has discussed fundamental elements of sensemaking as challenges to be considered in order to make robots explainable for ordinary people. Moreover, the main implications for the development of trustworthy relationships have been considered.

These factors, along with an analysis of existing models for explanatory interactions, provided groundwork for proposing a comprehensive framework to model explanatory interactions with social robots. At the core of this model are the contextual nature of explanations, the possibility of iterating them and the use of combined signals in order to maximize the chances of successful understanding. Nevertheless, the scarcity of long-term exposure to these novel technologies makes it difficult to precisely predict how human parties will adapt to explainable robots in terms of trust.

Moreover, the possibility that users fail to understand robots’ explanations despite repeated attempts is a fundamental limitation of any approach to explainable robots.

Finally, given its conceptual and theoretical nature, the main limitation of this article is the lack of a user study. Therefore, a continuation of this work will be to test how the proposed model for iterated and multimodal explanatory interactions influences the overall robot-user relationship, specifically how the examination phase can be calibrated through the use of combined signals as the interaction develops in order to improve users’ understanding and trust toward robots.

References

[1] E. Martinez-Martin and A. P. del Pobil, “Personal robot assistants for elderly care: an overview,” in *Personal Assistants: Emerging Computational Technologies*, A. Costa, V. Julian, and P. Novais, Eds., Springer, Cham, Switzerland, 2018.

[2] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, “Social robots for education: a review,” *Science Robotics*, vol. 3, no. 21, pp. 1–9, 2018.

[3] A. Tapus, M. J. Mataric, and B. Scassellati, “Socially assistive robotics [grand challenges of robotics],” *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 35–42, 2007.

[4] M. M. De Graaf, B. F. Malle, A. Dragan, and T. Ziemke, “Explainable robotic systems,” in *Proc. HRI'18 Companion*, ACM, 2018, Chicago, Illinois, USA, 2018, pp. 387–388.

[5] M. M. De Graaf and B. F. Malle, “How people explain action (and autonomous intelligent systems should too),” in *Proc. AAAI Fall Symposium Series*, AAAI, Arlington, Virginia, USA, 2017, pp. 19–26.

[6] M. Lomas, R. Chevalier, E. V. Cross, R. C. Garrett, J. Hoare, and M. Kopack, “Explaining robot actions,” in *Proc. HRI'12 Int.*

Conf., ACM, 2012, Boston, Massachusetts, USA, 2012, pp. 187–188.

[7] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[8] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[9] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda," in *Proc. CHI'18*, ACM, Montréal, QC, Canada, 2018, pp. 1–18.

[10] O. Biran and C. Cotton, "Explanation and justification in machine learning: a survey," *IJCAI-17 Workshop on Explainable AI* (*XAI*), vol. 8, no. 1, pp. 8–13, 2017.

[11] F. K. Došilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: a survey," in *Proc. 41st MIPRO Int. Conv.*, IEEE, 2018, Opatija, Croatia, 2018, pp. 0210–0215.

[12] S. Leigh Star, "This is not a boundary object: reflections on the origin of a concept," *Science, Technology, & Human Values*, vol. 35, no. 5, pp. 601–617, 2010.

[13] L. K. Berland and B. J. Reiser, "Making sense of argumentation and explanation," *Science Education*, vol. 93, no. 1, pp. 26–55, 2009.

[14] F. C. Keil, "Explanation and understanding," *Annu. Rev. Psychol.*, vol. 57, pp. 227–254, 2006.

[15] T. Lombrozo, "The structure and function of explanations," *Trends in Cognitive Sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[16] T. Lombrozo, "Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions," *Cognitive Psychology*, vol. 61, no. 4, pp. 303–332, 2010.

[17] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys* (*CSUR*), vol. 51, no. 5, pp. 1–42, 2018.

[18] F. Sado, C. K. Loo, M. Kerzel, and S. Wermter, "Explainable goal-driven agents and robots-a comprehensive review and new framework," *arXiv preprint arXiv:2004.09705*, 2020.

[19] T. B. Sheridan, "Human-robot interaction: status and challenges," *Human Factors*, vol. 58, no. 4, pp. 525–532, 2016.

[20] R. Campa, "The rise of social robots: a review of the recent literature," *Journal of Evolution and Technology*, vol. 26, no. 1, pp. 106–113, 2016.

[21] K. E. Weick, K. M. Sutcliffe, and D. Obstfeld, "Organizing and the process of sensemaking," *Organization Science*, vol. 16, no. 4, pp. 409–421, 2005.

[22] P. Andras, L. Esterle, M. Guckert, T. A. Han, P. R. Lewis, K. Milanovic, et al., "Trusting intelligent machines: deepening trust within socio-technical systems," *IEEE Technology and Society Magazine, IEEE*, vol. 37, no. 4, pp. 76–83, 2018.

[23] N. Luhmann, *Trust and Power*, Polity Press, Medford, Massachusetts, USA, 2017.

[24] E. Broadbent, K. Peri, N. Kerse, C. Jayawardena, I. Kuo, C. Datta, and B. MacDonald, "Robots in older people's homes to improve medication adherence and quality of life: a randomised cross-over trial," in *Proc. ICSR 2014*, Springer, Cham, Sydney, NSW, Australia, 2014, pp. 64–73.

[25] H. M. Gross, S. Mueller, C. Schroeter, M. Volkhardt, A. Scheidig, K. Debes, et al., "Robot companion for domestic health assistance: Implementation, test and case study under everyday conditions in private apartments," in *2015 IEEE/RSJ IROS*, IEEE, 2015, Hamburg, Germany, 2015, pp. 5992–5999.

[26] M. Vincze, W. Zagler, L. Lammer, A. Weiss, A. Huber, D. Fischinger, et al., "Towards a robot for supporting older people to stay longer independent at home," in *ISR/Robotik 2014*, VDE, 2014, Munich, Germany, 2014, pp. 1–7.

[27] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[28] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," in *Proc. IAAI'17 Conf.*, AAAI, 2017, San Francisco, California, USA, 2017, pp. 4762–4763.

[29] P. Langley, "Explainable agency in human-robot interaction," *Proc. AAAI Fall Symposium Series*, AAAI, 2016, Palo Alto, California, USA, 2016.

[30] R. K. Sheh, "Different XAI for different HRI," *Proc. AAAI Fall Symposium Series*, AAAI, 2017, Arlington, Virginia, USA, 2017, pp. 114–117.

[31] H. Hagras, "Toward human-understandable, explainable AI," *Computer*, vol. 51, no. 9, pp. 28–36, 2018.

[32] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[33] J. Sun, "Emotion recognition and expression in therapeutic social robot design," in *Proc. HAI'14*, ACM, 2014, Tsukuba, Japan, 2014, pp. 197–200.

[34] R. K. M. Sheh, "'Why did you do that?' Explainable intelligent robots," in *WS-17-10 AAAI'17*, AAAI, 2017, San Francisco, California, USA, 2017, pp. 628–634.

[35] A. Theodorou, R. H. Wortham, and J. J. Bryson, "Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots," in *AISB Workshop on Principles of Robotics*, Bath University Press, 2016, Sheffield, South Yorkshire, UK, 2016.

[36] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[37] D. J. Hilton, "Conversational processes and causal explanation," *Psychological Bulletin*, vol. 107, no. 1, pp. 65–81, 1990.

[38] T. Wang, "Gaining free or low-cost interpretability with interpretable partial substitute," in *Proc. MLR*, PMLR97, 2019, Long Beach, California, USA, 2019, pp. 6505–6514.

[39] D. Walton, "A dialogue system specification for explanation," *Synthese*, vol. 182, no. 3, pp. 349–374, 2011.

[40] F. C. Keil, "Folkscience: coarse interpretations of a complex reality," *Trends in Cognitive Sciences*, vol. 7, no. 8, pp. 368–373, 2003.

[41] E. Pronin, "The introspection illusion," *Advances in Experimental Social Psychology*, vol. 41, pp. 1–67, 2009.

[42] B. F. Malle, *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*, The MIT Press, Cambridge, Massachusetts, USA, 2006.

[43] S. Overgaard, "The problem of other minds: Wittgenstein's phenomenological perspective," *Phenomenology and the Cognitive Sciences*, vol. 5, no. 1, pp. 53–73, 2006.

[44] A. Avramides, *Other Minds*, Routledge, Abingdon, Oxfordshire, UK, 2000.

[45] B. F. Malle, J. M. Knobe, and S. E. Nelson, “Actor-observer asymmetries in explanations of behavior: new answers to an old question,” *Journal of Personality and Social Psychology*, vol. 93, no. 4, pp. 491–514, 2007.

[46] J. Tullio, A. K. Dey, J. Chalecki, and J. Fogarty, “How it works: a field study of non-technical users interacting with an intelligent system,” in *Proc. CHI'07 SIGCHI Conf. on Human Factors in Computing Systems*, ACM, 2007, San Jose, California, USA, 2007, pp. 31–40.

[47] F. J. C. Garcia, D. A. Robb, X. Liu, A. Laskov, P. Patron, and H. Hastie, “Explain yourself: a natural language interface for scrutable autonomous robots,” arXiv preprint arXiv:1803.02088, 2018.

[48] M. E. Pollack, S. Engberg, S. Thrun, L. Brown, J. T. Matthews, M. Montemerlo, et al., “Pearl: a mobile robotic assistant for the elderly,” in *AAAI Workshop on Automation as Eldercare*, AAAI, 2002, Edmonton, Alberta, Canada, vol. 2002.

[49] IBM Research Editorial Staff, “Cognitive machines assist independent living as we age,” https://www.ibm.com/blogs/research/2016/12/cognitive-assist [accessed: May 29 2020].

[50] S. Arsovski, H. Osipyan, A. D. Cheok, and I. O. Muniru, “Internet of speech: a conceptual model,” in *Proc. 3rd Int. Conf. on Creative Media, Design and Technology* (*REKA 2018*), Atlantis Press, 2018, Surakarta, Indonesia, 2018, pp. 359–363.

[51] B. F. Malle, “Attribution theories: how people make sense of behavior,” *Theories in Social Psychology*, vol. 23, pp. 72–95, 2011.

[52] R. Ramos-Garijo, M. Prats, P. J. Sanz, and A. P. Del Pobil, “An autonomous assistant robot for book manipulation in a library,” in *Proc. SMC'03*, IEEE, 2003, Washington, DC, USA, 2003, vol. 4, pp. 3912–3917.

[53] M. Mikawa, M. Yoshikawa, T. Tsujimura, and K. Tanaka, “Librarian robot controlled by mathematical aim model,” in *Proc. 2009 ICCAS-SICE*, IEEE, 2009, Fukuoka, Japan, 2009, pp. 1200–1205.

[54] M. S. Sreejith, S. Joy, A. Pal, B. S. Ryuh, and V. S. Kumar, “Conceptual design of a wi-fi and GPS based robotic library using an intelligent system,” *International Journal of Computer, Electrical, Automation, Control and Information Engineering, World Academy of Science, Engineering and Technology*, vol. 9, no. 12, pp. 2511–2515, 2015.

[55] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, “Overtrust of robots in emergency evacuation scenarios,” in *Proc. HRI'16*, IEEE, 2016, Christchurch, New Zealand, 2016, pp. 101–108.

[56] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, “Explainable agents and robots: Results from a systematic literature review,” in *Proc. AAMAS'19*, ACM, 2019, Montreal, QC, Canada, 2019, pp. 1078–1088.

[57] A. Cawsey, “User modelling in interactive explanations,” *User Modeling and User-Adapted Interaction*, vol. 3, no. 3, pp. 221–247, 1993.

[58] P. Madumal, T. Miller, F. Vetere, and L. Sonenberg, “Towards a grounded dialog model for explainable artificial intelligence,” arXiv preprint arXiv:1806.08055, 2018.

[59] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, “A grounded interaction protocol for explainable artificial intelligence,” in *Proc. AAMAS'19*, ACM, 2019, Montreal, QC, Canada, 2019, pp. 1033–1041.

[60] M. Niemelä, P. Heikkilä, and H. Lammi, “A social service robot in a shopping mall: expectations of the management, retailers and consumers,” in *Proc. HRI'17 Companion*, IEEE, 2017, Vienna, Austria, 2017, pp. 227–228.

[61] Y. Chen, F. Wu, W. Shuai, N. Wang, R. Chen, and X. Chen, “Kejia robot – an attractive shopping mall guider,” in *Proc. ICSR 2015*, Springer, Cham, 2015, Paris, France, 2015, pp. 145–154.

[62] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, “Evaluating the engagement with social robots,” *International Journal of Social Robotics*, vol. 7, no. 4, pp. 465–478, 2015.

[63] S. M. Fiore, T. J. Wiltshire, E. J. Lobato, F. G. Jentsch, W. H. Huang, and B. Axelrod, “Toward understanding social cues and signals in human-robot interaction: effects of robot gaze and proxemic behavior,” *Frontiers in Psychology*, vol. 4, art. 859, 2013.

[64] S. F. Warta, O. B. Newton, J. Song, A. Best, and S. M. Fiore, “Effects of social cues on social signals in human-robot interaction during a hallway navigation task,” in *Proc. HFES 2018*, SAGE Publications, 2018, Boston, Massachusetts, USA, 2018, vol. 62, no. 1, pp. 1128–1132.

[65] S. Thellman, A. Silvervarg, A. Gulz, and T. Ziemke, “Physical vs. virtual agent embodiment and effects on social interaction,” in *Proc. IVA 2016*, Springer, Cham, 2016, Los Angeles, California, USA, 2016, pp. 412–415.

[66] C. S. Peirce, *Pragmatism as a Principle and Method of Right Thinking: The 1903 Harvard Lectures on Pragmatism*, Suny Press, Albany, New York, USA, 1997.

[67] G. H. Harman, “The inference to the best explanation,” *The Philosophical Review*, vol. 74, no. 1, pp. 88–95, 1965.

[68] D. A. Wilkenfeld and T. Lombrozo, “Inference to the best explanation (IBE) versus explaining for the best inference (EBI),” *Science & Education*, vol. 24, no. 9-10, pp. 1059–1077, 2015.

[69] J. C. Zemla, S. Sloman, C. Bechlivanidis, and D. A. Lagnado, “Evaluating everyday explanations,” *Psychonomic Bulletin & Review*, vol. 24, no. 5, pp. 1488–1500, 2015.

[70] T. Lombrozo, “Simplicity and probability in causal explanation,” *Cognitive Psychology*, vol. 55, no. 3, pp. 232–257, 2007.

[71] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. K. Wong, “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, 2013, San Jose, California, USA, 2013, pp. 3–10.

[72] P. Thagard, “Explanatory coherence,” *Behavioral and Brain Sciences*, vol. 12, pp. 435–502, 1989.

[73] P. E. Dunne, S. Doutre, and T. Bench-Capon, “Discovering inconsistency through examination dialogues,” in *Proc. IJCAI'15*, Morgan Kaufmann Publishers Inc., 2005, San Francisco, California, USA, 2005, pp. 1680–1681.

[74] T. Umetani, S. Aoki, K. Akiyama, R. Mashimo, T. Kitamura, and A. Nadamoto, “Scalable component-based Manzai robots as automated funny content generators,” *Journal of Robotics and Mechatronics*, vol. 28, pp. 862–869, 2016.

[75] H. P. Grice, “Logic and conversation,” in *Speech Acts*, P. Cole, J. L. Morgan, Eds., Brill, Leiden, The Netherlands, 1975, pp. 41–58.

[76] D. Walton, “Examination dialogue: an argumentation framework for critically questioning an expert opinion,” *Journal of Pragmatics*, vol. 38, no. 5, pp. 745–777, 2006.

[77] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, “Deception detection using real-life trial data,” in *Proc. ICMI'15*, ACM, 2015, Seattle, Washington, USA, 2015, pp. 59–66.

[78] R. Wiseman, C. Watt, L. ten Brinke, S. Porter, S. L. Couper, and C. Rankin, “The eyes don’t have it: Lie detection and neuro-linguistic programming,” *PLoS One*, vol. 7, no. 7, 2012.

[79] T. Hellström and S. Bensch, “Understandable robots – what, why, and how,” *J. Behav. Robot.*, vol. 9, pp. 110–123, 2018.

[80] R. A. Engle, “Not channels but composite signals: speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations,” in *Proc. 20th Cognitive Science Society Conf.*, Lawrence Erlbaum Associates, 1998, Madison, Wisconsin, USA, 1998, pp. 321–326.

[81] D. Huk Park, L. Anne Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, et al., “Multimodal explanations: justifying decisions and pointing to the evidence,” in *Proc. CVPR'18*, IEEE, 2018, Salt Lake City, Utah, USA, 2018, pp. 8779–8788.