TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

Explainable artificial agents:

Considerations on trust, understanding, and the attribution of mental states

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Dr.rer.soc.oec.

by

MSc Guglielmo Papagni
Registration Number 11832421

to the Faculty of Mechanical and Industrial Engineering

at the TU Wien

Advisor: Univ.Prof.in Mag.a rer.soc.oec. Dr.in rer.soc.oec. Sabine Koeszegi

The dissertation has been reviewed by:

Univ.Prof.Dr. Tom Ziemke	Univ.Prof.Dr. Edoardo Datteri

Vienna, 16th December, 2022

Guglielmo Papagni

Declaration of Authorship

MSc Guglielmo Papagni
Theresianumgasse 27, 1040 Vienna, Austria

I hereby declare that I have written this Doctoral Thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work - including tables, maps and figures - which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.

Vienna, 16th December, 2022

Guglielmo Papagni

Acknowledgements

In the first place, I would like to express my gratitude to my supervisor Sabine Koeszegi, whom I thank for believing in me, as a researcher and as a person. This achievement would have not been possible without her knowledge, experience and continuous mentoring.
I am also sincerely thankful to my parents, for pushing and motivating me in the pursue of this goal and to Carina, for the inspiration she constantly provides and for sharing this journey with me. Likewise, my appreciation goes to Ludovica, Maddalena, Giacomo, Carlo, Anys, and Tommaso for being there for me with their precious advice throughout these four years and beyond.
Finally, I would like to thank my colleagues from both the Institute of Management Science and the TrustRobots Doctoral College for the insightful conversations and support with the various projects that make up this dissertation.

Abstract

The use of artificial agents (i.e., artificial intelligence and physical robots) is increasing in a wide range of application contexts, many of which already concern the daily lives of non-expert users. For artificial agents to be socially accepted, it is fundamental that users place calibrated trust (i.e., not too much, not too little) in them. In turn, this depends on several factors, which include artificial agents performance, accuracy and, importantly, understandability.

This dissertation addresses a set of challenges that need to be overcome to successfully make artificial agents explainable and understandable by non-expert users. To this end, as explanations represent a fundamental form of social communication which has been thoroughly studied by social sciences, the first challenge tackled by this dissertation is of multidisciplinary nature. Here, we aim to integrate findings from social sciences into the design of explainable artificial agents. In particular, drawing from Karl Weick's 'sensemaking theory', this dissertation proposes a model for explanatory interactions with artificial agents.

Furthermore, this dissertation identifies factors that influence trust development over time. Additionally, the beginning of an interaction and the occurrence of unexpected events are found to be the situations that most likely require artificial agents to provide explanations for. Accordingly, this dissertation reports the results of an experimental study on which these theoretical considerations are tested by means of a mixed methodology investigation. Our main findings concern explanations' positive role as a trust restoration strategy, as well as the influence of 'institutional' cues and individuals' personality traits (e.g., propensity to take risks) in determining trust development.

Finally, this dissertation discusses how explanations typically refer to either intentional (i.e., intentions, reasons etc.) or unintentional (i.e., accidental, natural etc.) factors. This is of particular relevance for artificial agents, as they do not possess the genuine mental states required by biological intentionality and yet people easily attribute such qualities to them. This dissertation states that the attribution of intentionality to artificial agents is ethical, as long as their artificial nature is manifest. However, at the same time, artificial agents should support users, by means of explanations, with adopting the most adequate interpretative framework for each situation.

Keywords: Explainability, Trust, Understandability, Explainable artificial agents, Attribution of intentionality

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The introduction of this doctoral dissertation is an adapted version of the book chapter "Challenges and solutions for trustworthy explainable robots", written by the author of the dissertation for the "Doctoral College Trust Robots" book project, edited by the TU Wien Academic Press.

1.0.1 Background and motivation

The use of robots and other AI-based systems (henceforth called artificial agents) is increasing in a number of fields. This growth comes with a number of intertwined challenges of different nature (e.g., technical, social, legal etc.) that have to be faced for artificial agents to be accepted and integrated into society. Among these challenges, in recent years the idea that artificial agents should be able to explain their inner workings, decisions and actions has emerged in academic and societal debates. The intrinsic opaqueness of the algorithmic decision-making processes (i.e., 'black-boxes') frequently employed represents the main reason for the growing interest regarding explainability. This inscrutability may negatively impact people's understanding of artificial agents' behaviors, decisions and recommendations. In turn, failing to understand and predict how they behave, will likely lead some users to misplace trust [11, 40]. In fact, as on the one hand definitions of trust emphasize how uncertainties, vulnerability and perception of risk represent key elements that may jeopardize trust [31, 2, 35], artificial agents' lack of predictability may increase people's perception of uncertainty, hence undermining trust. On the other hand, studies also show how people's initial (i.e., not mediated by experience) perception of technology is mostly guided by individual dispositions and institutional cues, which could as well lead users to over-trust such machines.
Explanations, intended as a social communication tool that people use to justify events and behaviors (particularly if unexpected), find meanings, transfer knowledge and satisfy curiosity can support understanding and trust calibration (i.e., avoiding over- and under-trust) by shedding light on reasons and causes behind specific behaviors and events

[21, 38, 40, 12]. The process of seeking and providing explanations has been extensively studied within disciplines such as philosophy, sociology, psychology [21, 40]. Combining findings from such disciplines with the need to integrate them into artificial agents' design lies at the heart of what can be labeled as the interdisciplinary challenge of explainability [1].

As artificial agents are likely to permeate society on different levels that affect people's everyday life even more, their decision-making processes and behavior should be understandable not only for machine learning and robotics experts, but also for the broader audience of domain experts (i.e., practitioners from fields where such technologies are applied) and, importantly, non-expert users. Each of these categories of users has different needs, interests and familiarity with the technology. Therefore, it is fundamental to understand and acknowledge the differences between different types of users to determine what desiderata and goals explainability should pursue in each context.

The category of domain experts concerns applications such as military operations (e.g., robots used for finding and removing mines, or for rescue tasks), exploration (e.g., in space, or in the oceans), for medical purposes. This implies that most of the users will have to undergo some sort of specific training to interact with the machines. An initial training facilitates the creation of an adequate mental model of artificial agents which, in turn, supports users' understanding and trust calibration.

The category of non-expert users, on the other hand, refers to those users who mostly have little to no previous experience with specific robotic and AI technologies. It includes contexts such as care-giving and education, activities with recreational purpose and, perhaps more importantly, interactions with artificial agents "in the wild" [48]. This comprises mostly 'first-time' interactions that occur in open and rather uncontrolled contexts (such as shopping malls, stations, museums). Here, users will likely not receive any form of training and will now know exactly what to expect from the artificial agents.

This dissertation is primarily concerned with tailoring explainability to the needs of non-expert users for, at least, three reasons. In the first place, non-expert users represent the vast majority of the public and a large share of these technologies is designed to interact with them on a daily base. Furthermore, in reason of the limited technical knowledge and agency in terms of manipulating such technologies, the category of non-expert users represents the most vulnerable one. Finally, while the interests and needs of specific groups of users might differ, in principle an explanation that can be understood by users without any technical expertise should be comprehensible also to more technologically accustomed ones.

One of the main issues with tailoring explanations to the needs of non-expert users lies in the fact that explainability is often treated as a data-driven, rather than goal-driven quality [49]. On the other hand, social sciences have put a great deal of effort into investigating how people explain events and behaviors to each other, particularly in terms causal connections, explanations' structure and qualities as well as communication strategies [32, 21, 22, 40]. Hence, we claim that artificial agents' design should integrate inputs from other disciplines and focus on developing the capacity to communicate decisions in terms that are easily graspable by a broad and untrained audience. Another problem

that requires more extensive investigation is that explanations are, by their very nature, incomplete approximations of the actual decision-making processes [27, 47, 59]. The fact that perfect explanations do not exist is even more problematic for AI and robotics research in light of the standardized, algorithmic and "coordinate-based" modalities of processing information typical of artificial agents [33]. This calls for the development of implementable solutions that maximize users' chances of successful understanding.

1.0.2 Research questions and objectives

Committed to a multidisciplinary approach, this dissertation aims to combine findings from social sciences regarding explanations with advancements in AI and robotics. to do so, it addresses some of the key challenges of making artificial agents explainable and understandable, specifically in the context of everyday interactions and with non-expert users, to supports them with placing adequate trust in these technologies. As it was previously noted, for artificial agents to successfully integrate in our society, it is fundamental that users place adequate (i.e., not too much, not too little) trust in them. To this end, researchers suggest that adequate trust calibration is mediated by, among other things, correct understanding of artificial agents' behavior. However, as it occurs in human-human interactions, this process is not always straightforward.
People seek and provide explanations precisely because others' behavior and decisions are not always easy to interpret. However, even when provided with explanations, people may still struggle to understand the explanations and, consequently, the causes or reasons conveyed through the explanations. Making artificial agents explainable poses a multidisciplinary challenge at the heart of which is the integration of social sciences' understanding of 'explanatory interactions' into the design of artificial agents. To this extent, we investigate core concepts from Karl Weick's sensemaking theory, an interpretative framework developed, within the context of organizational sciences, to understand how people attribute meanings to events and others' behavior, particularly when these are ambiguous or unclear and induce the perception of uncertainty. Researchers propose models grounded in social sciences that describe ideal 'explanatory interactions' with artificial agents. However, sensemaking theory and its key proposals have not been investigated in the same context despite their relevance. Therefore, in paper 1 we investigate this potential contribution by addressing a set of questions.
RQ1: Can the core findings of sensemaking theory apply to the development of explainable artificial agents? If so, how should "explanatory interactions" be modelled accordingly?

Trust is a fundamental aspect of human-human interactions. It is studied by a number of disciplines which provide different perspectives ranging from trust within organizations to interpersonal relationships and, in recent years, in humans' interactions with automation and technology [46, 31, 50]. Trust is often defined as a trustor's belief that a trustee will help them achieve specific goals. As such, it rests on the trustor's belief that the trustee will behave benevolently but, at the same time, it implies accepting the risks, uncertainties and vulnerability deriving from the possibility that trust is misplaced [31, 35]. Hence,

trust is typically low when the perception of potential risks and uncertainties is, for one reason or the other, strong. Furthermore, placing trust in others is not a stable process. Rather, it should be investigated as a dynamic phenomenon. Studies in fields related to robotics and AI show how trust plays a key role in the process of acceptance of such new technologies [63, 33]. Therefore, what factors can influence (and how) the perception of artificial agents as trustworthy has become a central research topic. To this extent, an increasing number of studies relate explainability to trust in that the latter may change depending on how predictable artificial agents are [2, 24]. Explanations can provide useful insights to justify and clarify artificial agents' decisions and actions, particularly when these occur unexpectedly, and hence may support trust calibration. However, given the dynamic nature of trust, exactly how explanations can influence trust calibration over time remains largely uninvestigated. To this extent, the question that we investigate, both theoretically and empirically (paper 2 and 3), is the following.
RQ2: When are explanations mostly needed to support users' trust calibration and understanding of artificial agents' behavior, and how should explanatory content be provided to maximize the chances of achieving these goals?

Several studies show how people easily attribute social and mentalistic traits, such as reasons and intentions, to machines, despite the algorithmic nature of their decision-making processes does not engender genuine mental states. Much of the work in this direction refers to or is directly inspired by Daniel Dennett's concept of the 'intentional stance'.
When people explain events and behavior to each other, they refer to either mental states to clarify the reasons for specific behaviors, or to other factors that do not depend on one's intentions, desires and goals such as natural, mechanical and accidental causes. Recalling the fact that artificial agents do not have genuine mental states, intuitively it should be enough for them to explain their behavior in 'unintentional' terms. However, given that people are trained to and feel comfortable with attributing intentions and other mental states to machines, a mentalistic framework, with explanations that refer to such mental states, may as well be the most appropriate, or at least a practical choice. According to Dennett, modern, seemingly intelligent machines have reached a degree of internal complexity that makes it difficult, if not impossible, to make sense of their behavior in mechanical terms. The only chance is to adopt the intentional stance, which means treating machines 'as if' they had intentions, reasons and other mental states [13, 14]. While some researchers argue in favor of an intentional framework for artificial agents' explanations, others suggest that, in certain cases, trying to make sense of their behavior from a mentalistic perspective may not be the best strategy as it could result in cognitive conflicts [61, **?**, 45, 43].
To this extent, John Searle notes that Dennett, as well as a significant share of the debate fail to address the substantial difference between treating something (e.g., a machine) 'as if' it had mental states and believing that something genuinely and intrinsically has mental states. Whether an artificial agent's behavior is treated as intentional or unintentional changes how it is explained. Despite the abundance of studies research on the attribution

of mental states to machines, the connection with artificial agents' explainability is rarely investigated in a thorough fashion. Likewise, research often focuses on the *de facto* attribution of mentalistic traits to machines, overlooking ethical considerations of the phenomenon. Therefore, papers 4 and 5 address some of these open questions.
RQ3: Given that people tend to easily attribute social and intelligent traits to artificial agents despite these do not possess the qualities required for genuinely intelligent social behavior, should artificial agents explain themselves resorting to mental states such as intentions, reasons etc., or only in terms that refer to mechanistic, natural and accidental factors? Is it ethical to treat machines 'as if' they had mental states? Is it possible to take advantage of the practical benefits (in terms of explanations) deriving from the adoption of the intentional stance without committing, as Dennett does, to behaviorism?

1.0.3 Methodology and papers' contribution

This dissertation provides an exploratory approach to the topic of explainable artificial agent. The contribution of this dissertation, which consists of five papers, is primarily conceptual. However, some of the main propositions concerning explanations as a strategy to support trust calibration have been empirically tested (paper 3). What each of these paper entails is hereby briefly described.

Paper 1 The first paper contributes to the understanding of 'explanatory human-agent interactions' from the perspective of Karl Weick's sensemaking theory. Specifically, the paper stresses that for explanations to be considered successful in supporting trust calibration, they must in the first place be understood by users. To this end, three issues intrinsic of explanatory interactions are addressed. Namely, that explanations are, by nature, incomplete approximations of the actual decision-making processes, that explanations' validity is contextual and, finally, that the design of explainable artificial agents must care for people's limited 'explanatory forms of understanding' and knowledge retention [26, 27]. To address these problems, this paper develops a model for explainable artificial agents that is informed by the following propositions of sensemaking theory:

- Sensemaking is often taken for granted.
- The activity of sensemaking should be lifted from the implicit and private to the explicit and public sphere.
- Meanings in social interactions result from negotiations between the concerned parties and in specific contexts.
- Every new interaction represents a new sensemaking negotiation.
- When trying to make sense of events, people seek plausible stories, rather than accurate ones.

As Berland notes, the "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" [4, p. 27]. Accordingly, there is no single model to describe a perfect explanatory interaction, and any attempt to modeling interactions with artificial agents must be suitable for their algorithmic information-processing units. Existing models suggest that an explanation request is triggered by an "anomaly detection" [58] or a "knowledge discrepancy" [37, 36]. This approach expresses the idea of explanations as isolated events, rather than, as a sensemaking-based perspective suggests, contextual instances (see points 3 and 4 in the list above). Cawsey states that artificial agents should not assume what users know. Rather, users should be treated as 'novices' and artificial agents should update their mental model of the users as the interaction unfolds [7]. Recalling how every new interaction represents a new sensemaking negotiation, this paper claims that artificial agents should provide explanations about their role and functionality at the beginning of every interaction. This is particularly relevant considering that many future interactions are likely to occur 'in the wild' [48], which means that users will likely have little to no idea of what a specific artificial agent can and cannot do.

In regard to the approximate nature of explanations, this paper proposes a goal-driven approach tied to the concept of explanations' plausibility. To this extent, Weick notes that sensemaking intended as a process, "is driven by plausibility rather than accuracy" [60, p. 415]. Building upon Peirce's work on abductive reasoning, Wilkenfeld and Lombrozo rework Harman's concept of 'inference for the best explanation' [20, 44, 62]. Specifically, they postulate that the purpose of explainability should be to provide the best understanding of the likely causes of an event, rather than the most accurate explanation possible.

From a sensemaking perspective, the process of building meanings is the result of a collaborative and contextual effort involving the two parties (i.e., the explainer and explainee) [60]. In terms of explanations' plausibility, this entails that the explainee must understand and agree with the explainer that a specific explanation is plausible and sound as it sheds light on an event's most likely causes (i.e., it is unlikely for someone to find something plausible without understanding it in the first place).

Finally, the idea that successful explanations are those that users understand is connected to the problem that sensemaking is often taken for granted. Rather, as Weick notes, the activity of sensemaking should be lifted from the implicit and private to the explicit and public sphere. Previous models either suggest that a sufficient criterion to assess understanding is that a user states they understood an explanation [37, 36], or that users should be questioned about their understanding [58]. This paper identifies and discusses two strategies that may support users' successful understanding. Respectively, these are explanations' 'multimodality' and 'interactivity'.

The former refers to artificial agents' possibility to convey explanatory information not only by means of natural language, but also through the 'combination' of multiple communication channels [18]. Depending on an artificial agent's embodiment, several techniques, such as graphical visualization, expressive bodily motion and speech are

available beside text-based communication and may improve explanations' quality and understandability [23, 25, 3].
Then, 'interactivity' refers to making explanatory interactions more human-like by approaching them as dialogues rather than 'single-shot' utterances. Importantly, 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') [26]. Several strategies are available to engender explanations' interactivity. Introducing 'nested argumentation dialogues' increases the interaction naturalness and allows users to engage in multilayered explanations in which they can drift from one question to another, for instance by asking for further details, in a back-and-forth manner [37, 36]. Even when explanations appear sound and plausible, they may still be grounded upon incorrect premises [58, 16, 29]. Implementing a dialectical shift in the form of an 'examination phase' allows users to search for inconsistencies, potential errors and, ultimately question explanations' truthfulness [16, 30, 57, 58]. While this property alone can improve explanations' quality, another possible use for an examination phase is to test the explainee's understanding of an explanation [58]. Leveraging on Weick's intuitions (see points 1 and 2 in the list above), this paper proposes an incremental and contextual approach, which implies that users' understanding may be questioned, but only proportionally to the time they can invest in the interaction.

Paper 2 This paper provides a conceptual analysis of the intertwining between trust and explainability aiming to determine when explanations are mostly useful in supporting trust calibration. To answer this question, we conducted a preliminary analysis of what trust in human-agent interaction is. Specifically, this analysis focused on factors that influence initial trust, as well as trust development over time, included violations and restoration strategies. To this extent, we found that initial trust results from a combination of both personal attitude toward technology and 'institutional cues' [54, 2]. The former is a consequence of the combination of several factors, such as cultural background, demographics, and personality traits [41, 8], 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 [17, 12], to skepticism and even forms of 'technophobia' [28].
The notion that trust partially depends on 'institutional cues' refers to the role played by 'third parties', such as private companies, developers, national and international institutions, 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.
Based primarily on 'institutional cues' and individual attitude, initial trust can be very high or low irrespective of artificial agents' actual performance (i.e., not calibrated). For this reason, and contrarily to the idea that an explanatory interaction begins necessarily with a 'knowledge discrepancy' or an 'anomaly detection' [58, 36], we emphasize the importance of artificial agents' initial explanations. In fact, when they 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 [2, 19].
A 'knowledge discrepancy' or anomalous and unpredictable behavior represents the other moment in an interaction when people seek out explanations [2, 40]. In other words, once users establish a mental model of an artificial agent based on prior interactions, this will be expected to perform actions within a certain range of possibilities. Within this range, the artificial agent's reliability will be progressively determined based on its performance and accuracy. As reliability and trustworthiness are proven over time, users may consolidate their positive mental model, so that explanations become superfluous if not even damaging [15]. At this point of an interaction, an artificial agent may still act unexpectedly or unpredictably. Such events, particularly if they turn out to be actual mistakes, become crystallized instances of the interactions and may cause a re-calibration of users' mental models. In similar situations, users' understanding of the agent's behavior is challenged and their acceptance and trust in the agent may be at stake [33, 10, 11, 40]. Whether it is to prevent the loss of trust (in case of a misunderstanding), or restore it after a mistake, we claim that, whenever required by the concerned users, artificial agents must provide reasons for their actions through explanations.

Paper 3 Based on the conceptual findings from paper 2, we designed a user study to test our propositions about the intertwining of explainability and trust. To this end, we mimicked using the Wizard of Oz methodology (WoZ) the functionalities of a personalized virtual learning assistant. Its customized features included personalized reminders, scheduling and interface as well as the choice of four different learning styles. The assistant's explainability and accuracy were manipulated (between-subjects variables), with the interaction time being treated as a within-subjects variable, resulting in four experimental conditions. Respectively, these were labeled 'correct with explanation', 'correct without explanation' 'faulty with explanation' and 'faulty without explanation'. In the study, participants had the task to interact with the assistant seven times over the course of seven work days. The goal of the learning assistant was to provide participants with recommendations in the form of chunks of text (i.e., abstracts) obtained from longer texts so that they could prepare for quizzes. Participants always had access to the full texts, but accepting the recommendations would save them time. In the 'faulty' conditions, the assistant would make a wrong recommendation at the fourth interaction. Participants in both groups 'with explanation' could access information about how the natural language processing (NLP) algorithms of the assistant worked at the beginning and throughout the study. Another explanation clarifying the reasons for the wrong recommendation was offered by the assistant to participants in the 'faulty with explanation' group after the wrong recommendation. After the end of the experiment, participants were debriefed about the purposes and nature (WoZ) of the study. Only those who completed the procedure received a compensation. Trust was measured by means of validated questionnaires at the beginning, throughout the study and at the end. Additionally, qualitative data was collected by means of semi-structured interviews (n=18) and a focus group.

We collected a total of 171 complete observations for the quantitative analysis. Among the main findings, we observed that, contrary to expectations, initial explanations about the assistant's functionality did not increase initial trust. To this extent, in line with the idea that 'institutional cues' contribute to shape people's perception of new technologies in terms of trust, the qualitative data suggested that the researchers' 'hidden authority' had a favorable impact on participants' perception of the assistant. Furthermore, in accordance with the experimental propositions, the assistant's wrong recommendation affected participants' trust negatively, as it was perceived as a trust breach. Nevertheless, qualitative data revealed that participants tended to be quite tolerant toward imperfect AI-based systems, as such technologies are not expected to always function perfectly. Importantly, trust restoration was significantly faster when the assistant 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 increases transparency and trustworthiness. While the findings did not support all our expectations, the role of explanations as a trust restoration strategy was validated. Furthermore, insights from the qualitative provided us with more nuanced interpretations of the dynamics of trust development.

Paper 4 According to Weick's sensemaking theory, finding meanings in the social context of everyday life entails bringing order to the chaotic stream of both intentional behaviors and unintentional events [60]. As this also extends to artificial agents' behavior, in Paper 4 we conducted a conceptual analysis on the attribution of intentionality to artificial agents and its implications. This work focuses prominently on Daniel Dennett's formulation of the 'intentional stance', primarily because a significant share of the recent debate is informed, when not directly inspired by his work.

To this extent, while acknowledging the importance and pragmatic usefulness of an idea such as that of the intentional stance, we note how it comes with one main problematic aspect. This, as pointed out by authors such as John Searle And Ned Block, is the intrinsic behaviorism of Dennett's formulations. In his interpretation, treating a machine 'as if' it had mental states is equal to thinking that the machine has genuine mental states. In other words, manifest behavior is the only significant aspect to consider, because alternatively one would have to accept the existence of unfathomable qualities (i.e., 'qualia') [**?**, 9]. According to Searle, the conflation of the two concepts of having genuine mental states and treating an artificial agent 'as if' it had mental states is the core problem in Dennett's formulation [52, 56].

On top of this considerations, we then investigate the origin of the phenomenon from a comparative perspective that extends to different forms of artificial agents' embodiment. While several studies address the attribution of mental states to specific groups of artificial agents (e.g., robots or virtual agents), a more inclusive perspective has often been overlooked. To this extent, we found that the phenomenon seems to depend on the

'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 [6, 34, 55, 43]. Consequently, as most people appear to lack a strategy for interacting specifically with sophisticated technologies, this socio-cognitive process is easily triggered when interacting with seemingly intelligent machines. Problematically, this mechanism seems to work even when a mentalistic interpretation is not the most adequate. Resting on the idea of 'minimally rational' behavior [42], from our comparative stance we identify in the apparent rationality of artificial agents' behavior, more than in their appearance, internal complexity or movements, the key aspect for users to adopt an intentional or mentalistic framework.
At the same time, when an artificial agent's behavior does not carry any rational meaning, attempting to understand 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 [61]. Recalling the distinction highlighted by Searle, it is precisely when an artificial agents' behavior does not appear rational (e.g., in case of a mistake) that one must be able to tell the difference between 'intrinsic intentionality' and 'observer-relative ascription of intentionality' [51, 52]. To avoid such cognitive conflicts, we sustain that users should be assisted with switching to a mechanistic interpretative framework. As it is further illustrated in Paper 5, explanations offer a tool to facilitate this process.
Finally, we discuss ethical implications of treating artificial agents 'as if' they had intentions and other mental states. Working with Danaher's notion of 'deception' [9] on use cases from the human-robot and human-computer interaction (respectively, HRI and HCI) literature, we argue that, in principle supporting the attribution of mental states to artificial agents is not necessarily a form of deception. To the contrary, given people's familiarity with this interpretative framework, it can be beneficial for interactions. However, upon the analysis of potential risks entailed by an unwarranted and uncontrolled ascription of mental states, we also identify artificial agents' transparency about their nature, functionality and behavior as a key requisite.

Paper 5 This paper further develops and translate some of the main insights from Paper 4 in terms of explainability for artificial agents and combines it with the overarching goal of ensuring users' understanding of artificial agents' behavior. Together, these perspectives call for explainable agents that resort, depending on the circumstances, either to mentalistic properties (i.e., reasons, intentions, desires, and beliefs), as well as causes of a different nature (e.g., mechanical, accidental, natural). As Bossi et al. note, "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)" [5, p. 1].
Prior to the discussion on explanations, we further investigate analyze the problematic relation between intrinsic, biological intentionality and the phenomenon of ascribing intentionality to artificial agents. To avoid the confusion that often stems from the overlapping of these two concepts, we resort an alternative, folk-psychological definition

according to which intentionality is not only the byproduct of biological evolution, but also a social construct that helps people understand, explain and predict other's behavior and, hence may facilitate social interactions. Intended as such intentionality does not necessarily imply consciousness or self-awareness [39, 10].
Although one may ask for an explanation out just out of curiosity [53, 1], explanations are typically requested when users' mental models of artificial agents are challenged by unpredictable behaviors. 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 reliance 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 artificial agent with whom they are interacting for an explanation. Some cases will force a complete perspective (i.e., framework) switch, while others will not.
This paper claims that artificial agents must be designed to support users, by means of explanations, in adopting the most appropriate interaction framework in any given context. This is especially the case for the early stages of extensive adoption of artificial agents in everyday contexts. Indeed, these times are most characterized by uncertainty in terms of both the adoption of and narratives built around these technologies. Guiding users toward the most appropriate framework could either mean that they are indeed already adopting the best one, and perhaps ask for an explanation out of curiosity (or because they are not sure about the reasons for the behavior), or they are not. In the first case, an artificial agent should provide confirmation to the user that they are already adopting the right framework. In the second case, the agent should support the transition from one interpretative framework to another.
Ultimately, given that meanings in social interactions are contextually negotiated (e.g., by means of explanations) between the concerned parties [60, 38], we argue that it is not only important to consider whether an artificial agent 'intended' to behave in a certain way, but also how the user perceives a specific behavior. Hence, we identify and discuss four scenarios that refer to whether a specific behavior was aligned with a robot's objective (i.e., intentional), as well as to how users may initially perceive it. Furthermore, we discuss, by means of literature-based use cases, what explanations should entail in each of the four cases, which are reported below.

- Intentional and (correctly) interpreted through an intentional framework;
- Unintentional and (erroneously) interpreted through an intentional framework;
- Intentional and (erroneously) interpreted through an unintentional framework;
- Unintentional and (correctly) interpreted through an unintentional framework;

Finally, as a limitation we acknowledge the technical difficulties that artificial agents might encounter when trying to identify (and explain) their own errors, as well as infer users' mental states and interpretation of the agent's behavior. At the same time, we also

note how specific approaches to explainability, such as 'interactivity' and 'multi-modality' may nevertheless increase users' chances to adopt the right framework.

1.0.4 Discussion and future work

This dissertation broadens and deepens the debate on explainability in robotics and AI by addressing several open questions and limitations in the existing literature. Specifically, this dissertation contributes to the understanding of explainability along two main trajectories that, while addressing specific aspects of the topic are deeply intertwined.
The first one is twofold and concerns the relation between artificial agents explainability and users' trust and understandanding on the one hand, and the role of social sciences in defining explainability on the other. The second addresses the relation between explainability and the phenomenon of attributing mental states to artificial agents.
With regard to the first aspect, we found that, if properly designed, explanations can prevent over- as well as under-trust. Particularly, this dissertation found that explanations may support trust trust restoration after a violation. Contrarily to expectations, another finding was that explanations are not always useful for initial trust calibration. However, as paper three discusses, the 'institutional cues' embodied in the figures of the researchers responsible for the study may have overridden the effect of explanations. Future work shall investigate more thoroughly such dynamics, isolate the effect of explanations, and that of institutional cues treating them as different variables.
Importantly, to realize their potential as a trust support strategy, it is crucial that explanations are understandable by users which, in turn, requires addressing two issues that are typically overlooked by research on AI and robotics. Respectively, these are explanations' approximate nature and contextual validity, and the need for human-friendly communication (i.e., quality- rather than data-driven explanation). As these challenges call for a deeper integration of insights from social sciences, this dissertation draws on the core properties of Karl Weick's 'sensemaking theory' to model explanatory interactions with artificial agents. Specifically, this dissertation proposes a sensemaking-based model that leverages on approaches to explainability, such as interactivity and multi-modality, to maximize the chances of successful understanding. Future work shall investigate the effectiveness of these techniques in different interaction contexts. Furthermore, as this dissertation emphasizes how complex and delicate the task of assessing one's understanding of an explanation is, further empirical validation of the available techniques is needed.
Another facet that this dissertation contributes to concerns the relation between explainability and the tendency to ascribe intentions and other mental states to AI and robots. While the latter represents a topic with a rich and long-lasting research history in itself, at the same time it is also embedded in the study of explainable artificial agents. In fact, whether an event (or behavior) is treated as intentional or unintentional influences how it is explained.
To this regard, this dissertation found that the attribution of mental states incarnates human brains' disposition towards a mentalistic and social interpretation of artificial agents' behavior. The mechanism is easily triggered whenever certain conditions are met and it has been proven helpful to make sense of artificial agents' behavior. For this

reason, this dissertation claims that it is in principle ethical to implement design features that induce the adoption of the mentalistic framework. However, a precondition that this dissertation identifies is that developers, private companies and whoever is responsible for the introduction to the public of artificial agents should always let users be aware of the artificial nature of the agents, to avoid not only cognitive conflicts, but also deception. Furthermore, while in many cases a mentalistic interpretation may facilitate interactions, other circumstances will require a different framework (e.g., a mechanistic one). To this regard, this dissertation proposes that, by means of explanations, artificial agents should guide users to the most appropriate framework for every circumstance. Since the adoption of the wrong interpretative framework may cause cognitive conflict on the side of the user, who may consequently fail to understand the causes behind the events being explained, guiding users to adopting the right framework may also support trust calibration. Future work shall test empirically whether explanations lead users to the right framework under different interaction circumstances, as well as how this affects users' trust in the agents.

CHAPTER 2

Paper1

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

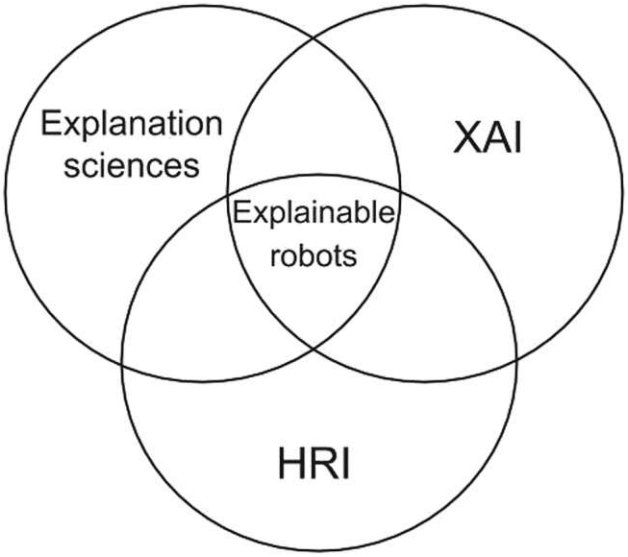


Figure 1: The interdisciplinary challenge of explainable robots.

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 sense-making 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].

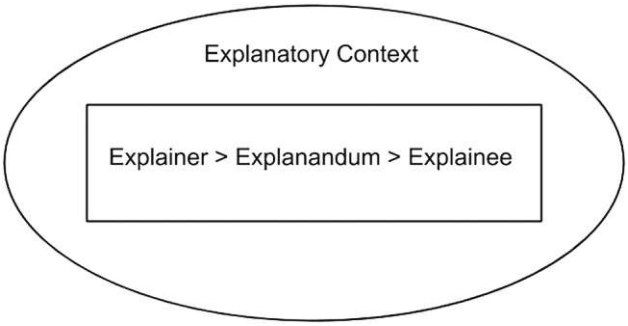


Figure 2: Explanations as contextual events.

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.

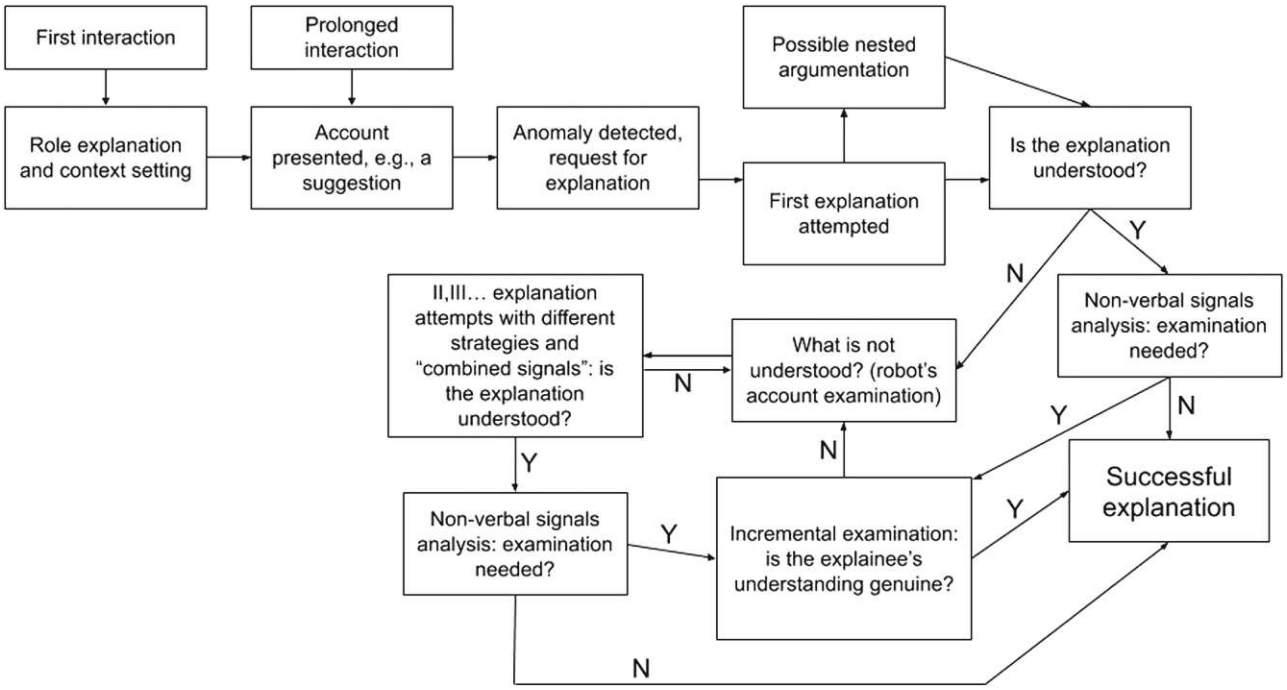


Figure 3: Explanatory dialogue model.

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.

CHAPTER 3

Paper2

Artificial Agents' Explainability to Support Trust: Considerations on Timing and Context

Guglielmo Papagni[1], Jesse de Pagter[1], Setareh Zafari[1], Michael Filzmoser[1], and Sabine T. Koeszegi[1]

[1]*Institute of Management Science, TU Wien, Vienna, Austria*

April, 2021

Acknowledgements Funding for this study was contributed by Mercedes-Benz AG. The authors would like to thank Alischa Rosenstein, Dimitra Theofanou-Fülbier and Joana Hois for their guidance and feedback throughout this project.

Abstract

Strategies for improving the explainability of artificial agents are a key approach to support the understandability of artificial agents' decision-making processes and their trustworthiness. However, since explanations are not inclined to standardization, finding solutions that fit the algorithmic-based decision-making processes of artificial agents poses a compelling challenge. This paper addresses the concept of trust in relation to complementary aspects that play a role in interpersonal and human-agent relationships, such as users' confidence and their perception of artificial agents' reliability. Particularly, this paper focuses on non-expert users' perspectives, since users with little technical knowledge are likely to benefit the most from "post-hoc", everyday explanations. Drawing upon the explainable AI and social sciences literature, this paper investigates how artificial agents's explainability and trust are interrelated at different stages of an interaction. Specifically, the possibility of implementing explainability as a trust building, trust maintenance and restoration strategy is investigated. To this extent, the paper identifies and discusses the intrinsic limits and fundamental features of explanations, such as structural qualities and communication strategies. Accordingly, this paper contributes to the debate by providing recommendations on how to maximize the effectiveness of explanations for supporting non-expert users' understanding and trust.

Keywords: Trust, Explainability, Artificial Intelligence, Explainable Artificial Agents

Table of Contents

Introduction

Trust is studied in a wide variety of disciplines, including social psychology, human factors, science and technology studies, and industrial organization, as understanding trust is relevant in many contexts. Each perspective implies a different interpretation of trust, ranging from interpersonal trust (Rotter, 1971; Simpson, 2007) and trust within organizations (Schoorman et al., 2007; Zaheer et al., 1998; Zucker, 1987) to trust across different levels of society such as between individuals and institutions and companies (Fulmer and Gelfand, 2012). In particular, increasing efforts have been made recently to investigate trust in the relationships between humans and machines. Despite multiple studies on trust in automation, conceptualizing trust over time and reliably modelling and measuring it remains a challenging issue Andras et al. (2018); Jacovi et al. (2021); Lockey et al. (2021). Likewise, there is a lack of a systematic perspective on how trust changes across different moments of an interaction and how it is influenced by different behaviors by artificial agents.

The main purpose of this paper is to provide a conceptual analysis of the connections between trust and explainability in the context of repeated human-agent interaction. Specifically, this paper aims to identify when explanations

are most useful as a trust support strategy and how they should be tailored accordingly. To meet our goal, we support our claims with use cases and examples from the literature on different types of artificial agents.

Importantly, this paper refers to the rather broad and inclusive term of 'artificial agents' to extend our considerations to different forms of artificial intelligence (AI) embodiment. Throughout the paper, we address specific types of agents such as virtual ones and physical robots by means of use cases to support our claims. Furthermore, the paper primarily focuses on interactions between non-expert users and artificial agents. We prioritize non-expert users because they represent the vast majority of the public. To this extent, someone who is a domain-expert in one field (e.g., a clinician or military personnel) will likely be a non-expert user in other situations. Perhaps more importantly, non-expert users' lack of knowledge about artificial agents' inner workings makes them a more vulnerable category (compared to domain experts and expert practitioners) (Lockey et al., 2021). Here, 'interaction' is generally intended as any social encounter between users and artificial agent, with particular attention being paid to 'long term' interactions.

Section 1 presents a discussion on the multifaceted concept of trust and those related to it such as reliability, confidence and familiarity in the context of day-to-day human-agent social relationships.

Importantly, as trust depends on users' capacity to predict an artificial agent's behavior (Jacovi et al., 2021), we identify the beginning of an interaction and when artificial agents behave unpredictably as the moments in which trust is more at stake (Andras et al., 2018). In the first case, users cannot resort on previous experience with a specific artificial agent to generate accurate predictions about the agent's future behavior In the second case, trust may be jeopardized by unexpected behaviors which could force users to adapt their mental models and, hence, their expectations and predictions about an agent's future behavior.

Particularly in relation to initial trust and acceptance of new technologies, the role played by 'third parties' responsible for the adoption and distribution of new technologies is further discussed (Coeckelbergh, 2018; Elia, 2009).

Explanations are often pointed at as an implementable strategy that may support trust. However, precisely why this is the case is often overlooked. Therefore, on top of the initial considerations on trust, Section 2 critically examines when and how explanations are most useful as a trust support strategy. We discuss what explanations are and present the idea of explanations' plausibility as a key quality that allows to match interactions' contextual affordances, artificial agents' availability and explanations' flexibility. We also identify 'approximation' and the possibility of being untruthful while being plausible as the main limits of explainability.

Building upon this, Section 3 focuses on explanations' communication strategies that support users' understanding while at the same time mitigating explainability' intrinsic limits. We identify in the combination of explanations' openness, questionability and multi-modality as a promising solution. At the end of Section 3, the main propositions developed throughout the paper are

graphically rendered ni the form of a model that describes the connections between explanations and trust. Section 4 concludes the study and discusses directions for future research.

1 Trusting artificial agents

Previous research on trust over time in human-agent interaction has primarily focused on identifying initial trust levels and potential determinants (Hancock et al., 2011; Salem et al., 2015). Short-term studies such as these are not necessarily capable of revealing (subtle) changes over time. Given the dynamic nature of trust (Holliday et al., 2016; Lyon et al., 2015), there is little understanding of how trust relationships with artificial agents can form and evolve over long periods of time. Few empirical studies investigate the fluidity of trust (Ho et al., 2017; van Maris et al., 2017). Recent long term studies (van Maris et al., 2017; Rossi et al., 2020) have found time to be an important factor influencing trust in repeated interactions between humans and robots. De Visser et al. (2020) presented a model for long-term trust calibration by providing techniques to mitigate over-trust and under-trust effects in robots. Taken together, these studies highlight the need to identify what aspects of a system's design and behavior determine the development of trust over longer periods of time. Upon the consideration of the dynamic and context-dependent nature of trust-based interactions (Holliday et al., 2016; Jacovi et al., 2021; Lee and See, 2004; Lyon et al., 2015), to meet our goal we first analyze what the literature recurrently highlights as the fundamental elements of trust in human-agent interaction that ought to be considered throughout the design and implementation phases of explainability strategies.

1.1 Fundamental features of trust

1.1.1 Risk, uncertainty, vulnerability

Andras et al. (2018) refer to the work of Luhmann (2018) and define trust towards artificial agents as the willingness to take risks amid uncertain conditions. Accordingly, Lockey et al. (2021) highlight how such conditions of risk and uncertainty requires people to take a 'leap of faith' and expose themselves to vulnerability. In line with these positions, (Lee and See, 2004, p. 51) define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability".

However, Lockey et al. (2021) clarify that one's willingness to face vulnerability must be motivated by positive expectations. In other words, trust's 'leap of faith' requires 'good reasons'. Otherwise, it would be a matter of 'blind fate' rather than trust.

1.1.2 Contextual nature of trust

Lee and See formulate of trust within a three-dimensional model so that trust is influenced by a person's knowledge (i.e., expectations and predictions) of what artificial agents are supposed to do (purpose), how they function (process), and their actual performance (Lee and See, 2004). In other words, people will grant trust to an artificial agents if they think or expect that the agent will perform according its 'purpose'.

Accordingly, Jacovi et al. (2021) argue that an AI model (i.e., the 'locus' of the decision-making processes) is trustworthy if it acts consistently according to specific 'contracts' the model or artificial agent entertains with a user. With artificial agents, these contracts may concern a wide variety of applications. For instance, if an AI model is employed as a recommender for an online streaming platform and does so successfully over time, then it can be considered trustworthy to the extent of providing users with suggestions about music, movies and so on.

The contractual or purpose-dependent essence of trust implies that users' expectations, predictions and willingness to grant trust should be confined within such specific boundaries. Holliday et al. (2016) similarly argue that people may contextually and contractually trust other agents in some regards, but such trust is not necessarily 'transferable' to other contexts.

1.1.3 Trust, reliability, and confidence

The contractual nature of trust-based interactions has an important timing-related component. To this extent, one element of the formulation of 'trustworthiness' by Jacovi et al. (2021) needs further discussion. The authors mention that a model is trustworthy if it acts 'consistently', which implies stable performance over time (the third element in Lee and See (2004)'s model). This recalls definitions of reliability, a term often associated with trust and trustworthiness. In fact, reliability can be defined as an artificial agent's capacity to achieve a specific goal in accordance with its purpose (Fossa, 2019; Lee and See, 2004). Reliability, intended as the 'capacity to act consistently', emerges as a quality that can be inferred only on the basis of past performance (O'neill, 2002).

Confidence, intended as the belief that a certain event will occur as expected, represents the counterpart (on the users' side) of reliability. As such, it is based on high familiarity and requires no explicit decision-making (Pieters, 2011). If an artificial agent proves to be reliable as it acts consistently in accordance with its purpose, people become confident about how the agent will behave in the future and will not necessarily have to explicitly assess its trustworthiness at each interaction. Once the agent's reliability has been established based on positive experiences, the perception of risks decreases. In other words, one becomes confident in the system's competence to fulfill its purposes (Gefen, 2000; Luhmann, 2000).

However, if there is no record of past performance, one cannot directly infer an artificial agent's reliability. One can only 'choose' to believe, that their

expectations and predictions about the system's future performance are accurate. In fact, unlike confidence, trust implies a decision-making process and the commitment to the accuracy of future performance (O'neill, 2002; Pieters, 2011; Taddeo and Floridi, 2011).

When people engage in an interaction with an artificial agent for the first time, they lack what Mollering (2006) defines as the 'routinary' aspect of trustworthy relationships. In the absence of the routinary and predictability aspects, trust implies the awareness that one's commitment might be wrongly placed (Pieters, 2011). However, if users' willingness to grant trust to an artificial agents is not supposed to be based on 'blind fate', their beliefs and expectations about how the agent will perform in the future need to be grounded on something else than the past performance record.

To this extent, several authors suggest that initial trust is primarily established upon individual dispositions and/or 'institutional cues' (Andras et al., 2018; Siau and Wang, 2018) and that, as interactions proceed, this initial attitude may be discredited or consolidated (Holliday et al., 2016; Lyon et al., 2015).

1.2 (Initial) trust establishment

A potential issue emerges here that concerns initial trust. In fact, on the one hand, individual dispositions towards technologies (especially new ones) are not always positive. On the other hand, institutions may operate as initial 'trustworthiness proxies', but the process is not always linear.

Concerning individual dispositions, various factors may contribute shaping users' initial attitude towards technology. Such dispositions may tend towards either a negative or an overconfident view on technology. These result in a wide variety of reactions that range from skepticism in the form of general suspicion, pessimism or even 'technophobia' and 'neo-luddism' (Kerschner and Ehlers, 2016), to high expectations about new technologies (De Visser et al., 2020; Dzindolet et al., 2003), opinions based on subjective norms (Li et al., 2008), age and gender differences (Morris and Venkatesh, 2000; Venkatesh et al., 2000), and cultural and social background (Im et al., 2011).

Each of these factors alone or combined with others has the potential to undermine the acceptance of new technologies before they have the chance to prove their trustworthiness.

Then, regarding institutions' role in promoting the adoption of new technologies like artificial agents, the reliability of the entity - or set of entities - that introduce such technologies may work as a 'proxy' that guarantee the agents' trustworthiness.

Trust towards these 'third parties' might by influenced by their reputation and users may consequently extend trust to the newly introduced technologies as the result of a conscious or subconscious choice. The reliability of these entities may guarantee that the new technology will perform according to 'agreed-upon quality standards' that such third parties respected up to that point.

The idea of transferring the burden of initial trust to a third party is embedded in the concept of a shared sense of moral trust, i.e., the idea that the entity will behave with integrity and benevolence rather than in a harmful or duplicitous way towards those who trust it (Elia, 2009; Lankton et al., 2015; Pu and Chen, 2007; Sood, 2018). However, such influence might not suffice to convince people (e.g., technology-averse) of the 'benevolence' and reliability of a specific new technology.

To the contrary, a scandal or particular ethical concerns around a certain product by a company may result in a loss of trust towards the company itself. This has recently been the case with `Google Duplex`, an autonomous voice assistant, capable of (among other things) booking appointments. One peculiarity of `Duplex` is the close resemblance to a human voice, made possible by the implementation of features such as 'speech disfluencies', brief interruptions that people typically fill with noises like 'um' or 'ah' (O'Leary, 2019). The implementation of similar design features that allow `Google Duplex` to pose as a human, without users necessarily knowing it triggered ethical critiques and trust-related issues concerning both the specific product and, more generally, `Google`'s intentions.

Generally, these concerns about third parties' attitudes towards the public motivate the claim that companies and corporations should take action to implement or further improve their policies towards transparency and accountability with respect to new technologies. Corporations and commercial entities "need not express their concern for transparency in terms of stakeholders' rights, but they must care about those rights" (Elia, 2009, p. 152). Such a form of distributed responsibility (or lack thereof) for artificial agents' transparency is what we identify as a trust-enabling or trust-disabling factor, which has repercussions for interaction between users and artificial agents. In other words, a fair distribution of responsibility should represent a *conditio sine qua non* for end-users to build trust-based interactions with artificial agents.

1.2.1 Artificial agents' opaque processes

If trust is the result of a decision about predictability and expectations, then it is fundamental for users to understand why artificial agents behave the way they do. Several authors agree that understanding artificial agents' decision-making is fundamental for people to develop trust towards them (De Graaf and Malle, 2017; de Graaf et al., 2018; Lomas et al., 2012; Riedl, 2019). This aspect calls for the consideration of another element particular to artificial agents, that has the potential to jeopardize users' trust. In Lee and See's model this is the 'process' dimension, or how an artificial agent actually functions internally (Lee and See, 2004).

To understand the issue intrinsic of the 'process' dimension, a distinction between artificial agents and other forms of automation is needed. In the latter case a system's behavior is pre-programmed and its performance is limited to specific sub-sets of actions that the system is designed to perform. Instead, the former can be defined as having 'agentic' capabilities, which enable them to

respond to situations that are not pre-programmed or anticipated in their design (Zafari and Koeszegi, 2018). More and more, a large share of what can be termed agentic capability is made possible by the algorithmic information processing underlying decision-making processes. Generally speaking, the efficiency and adaptability of such processes improve as systems grow more complex. Particularly for artificial agents that are powered by deep learning algorithms which generate the so-called 'black-box models', their decision-making processes are becoming progressively more inscrutable (Adadi and Berrada, 2018). While this is primarily the case for laypeople and domain experts, i.e., professionals and practitioners who work in the fields where AI is applied (Ferreira and Monteiro, 2020; Preece et al., 2018), expert practitioners such as programmers and developers are also affected (Kaur et al., 2020).

It is precisely this complexity that poses a major obstacle to non-expert users' understanding and sense-making processes and, hence to trust (Papagni and Koeszegi, 2020). Recalling Lee and See's model, while the quality of the performance generally improves thanks to the use of opaque models, people's knowledge and understanding of how artificial agents function internally decreases. However, if artificial agents prove to be reliable according to their purpose, users will likely grow confident and may not question how the decision-making processes actually work. This is not to say that understanding is not important when artificial agents perform well and consistently and users' confidence levels are high. It simply means that as long as artificial agents behave according to users' expectations and predictions, users will less likely question the artificial agents' reliability.

1.2.2 Unexpected events and trust violations

Even after artificial agents prove contractually reliable, users' confidence may still be affected and compromised forcing them to re-calibrate their expectations when artificial agents' behave unpredictably (Andras et al., 2018; Miller, 2019). The mismatch between users' expectations and artificial agents' actual behavior will likely result in a lack of understanding which, in turn, may negatively affect trust (Miller, 2019). In such cases, an artificial agent's past performance may not be a sufficient guarantee for levels of trust to remain high. If users do not understand artificial agents' behavior, this might be simply because the reasons behind such behavior are not immediately obvious.

However, if said behavior turns out to be a mistake, trust will be particularly at stake (Elangovan et al., 2007). Robinette et al. (2017) conducted a study in which participants were given the possibility to follow a robot's guidance to exit a risky situation. Their results show a significant decrease in self-reported trust when the robot failed the task, compared to when it performed successfully. Additionally, participants who experienced the failure were less prone to follow the robot's guidance in later interactions. Since autonomous systems are not perfect, a trust restoration strategy seems to represent a more viable solution compared to relying on perfectly accurate performance.

To summarize, our initial analysis showed how trust implies the expectation

that an agent will perform with consistence in regard to its purpose. At the same time, it always implies accepting risks and uncertainties and the the resulting vulnerability. It also emerged how trust is mostly at stake at the beginning of an interaction and when an artificial agent behaves unexpectedly. This is because initial trust (or lack thereof) depends on individuals' attitude and institutional players (such as commercial companies, legislators etc.), rather than on the expectations deriving from an artificial agent's actual capabilities. Then, if an agent behaves unexpectedly, this may cause users to fail understanding and, consequently, re-calibrate their expectations, possibly jeopardizing their trust.

The next section argues that the implementation of explainability may not only support users' understanding of artificial agents' actions and inner workings, but also support initial trust establishment as well as prevent, or at least mitigate trust losses in the context of repeated interactions.

2 Explainable artificial agents

Calls for increased transparency have been a central concern for several regulatory organs (Goodman and Flaxman, 2017; Gunning, 2017; Gunning and Aha, 2019; Hleg, 2019). Making artificial agents explainable is one possibility to achieve 'transparency' and 'interpretability'. Interpretability itself represents a controversial 'umbrella term' (Lipton and Steinhardt, 2018). Researchers tend to group the available approaches into two main categories: direct interpretability and post-hoc interpretability, also known as 'explainability' (Hagras, 2018; Lipton, 2016; Molnar, 2020).

As direct interpretability is a quality that few models feature (e.g., linear models such as decision trees), here we will focus only on post-hoc generated explanation. This represents the primary approach to make 'black-box' models, such as deep neural networks, interpretable (Lipton, 2016; Molnar, 2020). However, few important considerations emerge from the debate over different approaches to interpretability that must be taken into account. Post-hoc explanations are only approximations of the actual decision-making processes and require a second, simpler model to clarify how inputs are processed into outputs (Wang, 2019). In turn, this makes explanations potentially unreliable and open to manipulations which may hide biases to the advantage, for instance, of the proprietary companies that own the rights of use of specific algorithms (Rudin, 2018).

'Hybrid interpretability' represents a promising solution that combines the strengths of the other two approaches. Unlike post-hoc interpretability, where a linear model is used as the explainer (Wang, 2019), hybrid interpretability features linear models in a 'ante-hoc' fashion. Specifically, this entails replacing the black-box model with a more transparent linear one and test whether it can produce comparatively accurate predictions with a subset of input data. If this is not the case, the black-box model is employed together with its explainer (Wang and Lin, 2021). This implies that in those cases which require the use of black-box models, the chances of untruthful or biased explanations persist. Section

3 describes how making explanations 'questionable' and 'interactive' may help coping with this issue and maximize the chances of successful explanations.

2.1 Explanations as trust support strategy

It is often reported how explanations may be useful to support trust towards artificial agents, particularly due to the opaqueness of their decision-making processes. Without explanations, people may struggle to build accurate mental models of artificial agents (Holliday et al., 2016) and to understand how decisions and predictions are generated (De Graaf and Malle, 2017; de Graaf et al., 2018; Lomas et al., 2012). However, exactly how explanations support trust is often not discussed in detail. To better understand this point, we shall first discuss what explanations are.

What constitutes a 'proper' explanation is an open question. 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" (Berland and Reiser, 2009, p. 27). Miller reports Lewis' definition that "to explain an event is to provide some information about its causal history. In an act of explaining, someone who is in possession of some information about the causal history of some event — *explanatory information*, I shall call it — tries to convey it to someone else" (Lewis, 1986, p. 99) in (Miller, 2019) (italic in the original version).

Furthermore, the informative content of explanations (i.e., the 'explanandum') can be of either 'scientific' or 'everyday' type. Both concern events' 'causal histories', and subsets of causes are selected to generate explanations (Hesslow, 1988; Hilton et al., 2010), but the former type refers to scientific connections of various points in an event's causal chain, while the latter aims to clarify "why particular facts (events, properties, decisions, etc.) occurred" (Miller, 2019, p. 5). As this paper focuses primarily on non-expert users' interactions with artificial agents, everyday explanations are more relevant for our purposes. Everyday explanations are forms of social communication which, through different means (e.g., textual, visual etc.) aim at transferring knowledge (Hilton, 1990) and fill in information asymmetries between one or more 'explainers' and one or more 'explainees' (Malle et al., 2007). By means of explanations, people persuade each other and influence each others' impressions and opinions (Malle, 2011). Explanatory information is often 'contrastive', meaning that people mostly ask why events and actions occur in certain ways rather than in others (Miller, 2019). While explanations that answer 'why-questions' are fundamental to justify artificial agents' decisions, explanations to 'how-questions' are central for transparency as they help understand the processes that bring artificial agents to specific decisions (Pieters, 2011).

For knowledge transfers to be successful, it is important that explanations are understood which, in turn, implies their coherence both internally and with the explainee's beliefs (Lombrozo, 2007; Thagard, 1989). Here, it emerges how explanations may be helpful for supporting users' trust towards artificial agents as they allow a transfer of knowledge about the otherwise opaque artificial agents'

decision making processes. We reported how standardization is not one of the strengths of explainability (Berland and Reiser, 2009). However, this entails that explanations are open to potential customization. As autonomous agents increase their presence in numerous aspects of daily life, they will likely interact with very diverse types of users (Hois et al., 2019; Mohseni et al., 2018). Accordingly, each context of interaction will tend to privilege certain specific qualities over others.

For instance, in some contexts simplicity, accompanied by a low level of technicality may be desirable explanations (Cawsey, 1993; Lombrozo, 2007; Zemla et al., 2017). This could be the case with online recommending systems such as those featured by streaming platforms or news websites. A rather unusual suggestion on what to watch, read, or listen to may trigger users' curiosity. A similar event would likely be considered as a low-stake case, as one could simply decide to skip the recommendation. However, studies show that even in such rather low-stake situations users benefit from explanations in terms of perception of the system's performance and trustworthiness (Shin, 2021). Therefore, an explanation in a similar case should be rather simple and quick and, for instance, refer to feature of the suggested movie or song that closely match previous users' choices.

Then, other situations in which the consequences at stake are significant may require explanations to be complete and spare no details, even if their internal complexity increases (Kulesza et al., 2013; Zemla et al., 2017). For instance, if algorithms are employed to compute loan requests or job applications, explanations for rejected requests should be rather extensive and exhaustive. They may, for instance, show how the process was not internally biased by forms of discrimination that have nothing to do with applicants' merits (Bellamy et al., 2018). Such discrimination types can follow nuanced paths and be difficult to detect but, when exposed, they can undermine the trustworthiness of whole processes. Consequently, if specific groups or communities (e.g., in terms of ethnicity or gender (Zou and Schiebinger, 2018) become the target of discriminatory AI-based decision-making processes due to underlying biases, members of these groups may develop systematic distrust towards AI-based technologies. In turn, the resulting lack of data including these discriminated groups in training data sets could further increase inequalities in automated decision-making processes, creating a vicious circle. In light of the context-dependence of what qualities explanations should have, we propose tailoring explanations according to the plausibility principle to maximize the benefits of explanations' flexibility and personalization options.

2.1.1 Explanation plausibility

In the field of explanation science, the relevance of explanations' plausibility can be found in the pioneering work on abductive reasoning by Peirce (1997). According to the author, explaining something is better described in terms of abductive reasoning as opposed to other cognitive process such as induction and deduction. Abductive reasoning involves proceeding from effects to causes

(like inductive reasoning). However, in deriving hypotheses to explain events, abductive reasoning assumes that something 'might be', rather than simply 'actually is' (Peirce, 1997).

Abductive reasoning has been interpreted as a process of 'inference to the best explanation' (Harman, 1965), which implies that explanations (ideally the best possible) are considered as the product of inferring processes. Perhaps more importantly for our purposes, Wilkenfeld and Lombrozo (2015) reformulate the concept emphasizing the processual nature of providing explanations. Intended as the process rather than a product, explaining something aims to trigger 'the best inference' possible. Importantly, this translates into the idea that people do not necessarily seek 'the true story'. They rather seek out plausible stories that can help them grasp the likely causes of an event (Weick et al., 2005).

So interpreted, abductive reasoning offers a reading in which plausibility emerges as a key criterion for selecting a subset of causes that could explain an event, where the explanatory power of an explanation is not a default quality but rather co-constructed by the parties. In this sense, plausibility implies that the soundness of the causes suggested to explain an event is determined by both the explainer, who offers the explanation, and the explainee, who evaluates it as sound. Furthermore, plausibility as a joint achievement represents the contextual sum of several explanation qualities that researchers identify as desirable.

A study from Wiegand et al. (2019) provides an example of how to tailor artificial agents' explanations according to the plausibility principle in the context of autonomous vehicles in a simulated environment. Specifically, they discuss how a self-driving car's explanations may be designed by combining inputs, in terms of mental model of the vehicle, from both experts and non-expert users (i.e., the typical 'passenger' of autonomous vehicles). The result is a 'target' mental model made out of those shared features that are identified as fundamental. This target mental model serves as a baseline upon which the cars explanations ought to be built. Interestingly, the authors also specify that, since participants in the study never had to take over the steering wheel, there was no timing limitation for interpreting the car's explanations.

Two problematic considerations need to be addressed in relation to plausibility. Some authors note that, in principle, an explanation might appear plausible but nevertheless be based on incorrect premises (Dunne et al., 2005; Lakkaraju and Bastani, 2020; Walton, 2011). When explanations are generated based on false beliefs, they can reinforce inaccuracies (Lombrozo, 2006) and thus incorrect mental models. This is the case when the plausibility of an explanation does not match its truthfulness. Furthermore, interpreting plausibility as 'explaining for the best inference' means looking at plausibility as a dynamic concept that is contextually negotiated between the interested parties at each explanatory interaction, rather than a fixed property. This may represent an issue, considering artificial agents' 'coordinate-based' reasoning (Lomas et al., 2012). Section 3 discusses explanations' 'interactivity' and 'questionability' as implementable strategies to cope with both issues.

2.2 Explanations' timing

We previously noted how, in the context of long-term interactions, trust in artificial agents is more likely to require direct support in two specific moments: in the case of a first interaction and when something unexpected happens.

2.2.1 Explanations to support initial trust

Andras et al. stress that explanations can support both the creation of appropriate mental models and initial trust when there is no previous experience as they may reduce the perception of risks and uncertainties (Andras et al., 2018). Accordingly, Cawsey (1993) suggests that, at the beginning of an explanatory interaction, explainees should be treated as 'novices'. This implies that artificial agents involved in the interaction should not infer what kind of mental model (of the agents) users already possess. Users should rather be supported, by means of explanations, to create an initial mental model of the artificial agents. Only as the interaction progresses, the artificial agents may infer what users know (Cawsey, 1993). Therefore, 'initial' explanations should primarily comprise information about the purpose of an artificial agent in a given interaction context.

This aspect is even more significant considering that a growing number of interactions with artificial agents will occur 'in the wild'. This includes interactions with artificial agents in 'uncontrolled' environments, as opposed to controlled ones were users are introduced and briefed about the agents' purpose and functionality. For instance, social robots are being tested as shopping mall assistants, with purposes that include entertaining customers, providing them with recommendations and guidance, and supporting retailers (Chen et al., 2015; Niemelä et al., 2017). If one such robot were to approach new potential customers, these would likely not know the robot's purpose. Initial explanations tailored to answer questions such as "what is the purpose of the robot/of interacting with it, why and to which extent should I trust it?" would help users establish a more accurate initial mental model, better understand how the robot can be helpful and, consequently, deciding whether to follow its suggestions and guidance.

2.2.2 Trust maintenance, calibration and restoration

Existing models of explanatory interactions with artificial agents identify an 'anomaly detection' or 'knowledge discrepancy' (on the part of the explainee) in the explainer's account as the trigger for explanation requests (Madumal et al., 2018, 2019; Walton, 2011). Unpredictable events represent a perfect example of such anomalies, as they 'abnormally' diverge from the expected course of events (Hilton and Slugoski, 1986; Kahneman and Tversky, 1981). Particularly, if these unexpected events turn out to be mistakes or errors, as these become part of the artificial agent's performance record, its reliability and trustworthiness may be shaken as users may be forced to re-calibrate their initially established mental model of the agent (Elangovan et al., 2007; Robinette et al., 2017). In other

words, after an unexpected event users may be wondering why did the agent behave in such a way and whether it makes sense to further grant trust to it. However, unexpected actions and behavior are not necessarily errors. It could as well be that the actual reasons behind the agent's behavior are not immediately obvious to the users, while still being plausible (Papagni and Koeszegi, 2021). Without explanations, it may nevertheless be difficult for users to determine whether unexpected behavior is the result of an actual mistake or just of a 'mental model mismatch'.

In similar circumstances explanations help not only restore, but also maintain trust. Conversely, it is likely that in 'in-between situations', i.e., when an artificial agent's performance is accurate, users will not need to update their mental models and the agent's trustworthiness and reliability will consolidate. Here, and more generally when users feel confident with the interaction tasks, explanations might be superfluous (Doshi-Velez and Kim, 2017). To this extent, Woodcock et al. (2021) conducted a study with non-expert users who had to evaluate explanations for diagnosis provided by an artificial intelligence-driven symptom checker. Their results suggest that high familiarity with specific diseases (e.g., migraine) may reduce explanations' positive effect on trust. However, explanations are ultimately not only useful to justify decisions, but may also satisfy users' curiosity and even help them learn and discovery something new (Adadi and Berrada, 2018). Therefore, in principle, artificial agents should always make them available to users and display them upon request.

Additionally, explanations may prevent users from overtrusting artificial agents (Lockey et al., 2021). In fact, some people tend to either have high expectation of technology (automation bias) (De Visser et al., 2020; Dzindolet et al., 2003) or to misjudge the risks implied by artificial agents' actions (Robinette et al., 2016; Wagner et al., 2018). However, at the same time skepticism towards technology is also a relatively common phenomenon (Kerschner and Ehlers, 2016). By providing users with a calibrated framework within which to interpret their behavior, artificial agents' explanations support users in both developing more accurate mental models and expectations as well as mitigating individuals' more extreme and, at times unmotivated, dispositions. Conversely, if an artificial agent does not perform very effectively over time, it is quite understandable for people to lose their trust until proven otherwise.

Other strategies exist to restore trust that, like explainability, can be implemented in human-agent relationships as well (Quinn et al., 2017). These capture both short-term and long-term perspectives and include denial, apologies, compensation and restructuring relationships (Lewicki and Brinsfield, 2017). However, we consider explainability a more appropriate strategy for at least two reasons. As discussed above, explanations have the twofold function of supporting both initial trust as well as trust maintenance and restoration, and should therefore be preferred over the application of multiple strategies. Furthermore, while alternative strategies such as apologizing or offering compensation might, in principle, help to regain trust, they do not offer much room for understanding the reasons behind specific events and actions. To this extent, fixing issues (e.g., bugs) that cause artificial agents' errors and the consequent improvement are

two of the main desiderata of explainability (Adadi and Berrada, 2018).

Before discussing how artificial agents may communicate explanations according to their specific affordances, we shall summarize the main points about explanations as trust support strategies. As it is graphically rendered in 1, explanations at the beginning of an interaction may support trust establishment by informing users about an artificial agent's role and introducing them to the interaction. Then, during the normal course of interactions artificial agents should be able to prove reliable, as long as they perform consistently in accordance with their purpose. However, users may be curious throughout an interaction about certain behaviors. Hence, even when an artificial agent performs consistently, it should be able to provide explanations upon clarification request and as a strategy to maintain trust. Finally, it may be that certain actions occur unexpectedly. To prevent (or mitigate, in case of a mistake) trust losses, artificial agents should be able to explain the reasons why things happened a certain way.

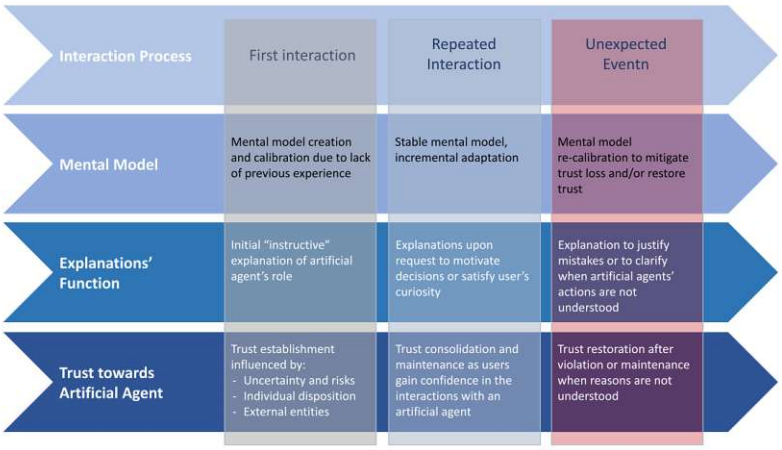


Figure 1: Graphic visualization of explanations as trust support strategy, throughout repeated interactions

3 Communicating explanations

We previously noted how explanations come with at least two major limitations. On the one hand, they only represent approximations of the actual decision-making processes. As such, they might appear plausible but nevertheless be based on incorrect premises, hide biases and be manipulated (Dunne et al., 2005; Lakkaraju and Bastani, 2020; Rudin, 2018; Walton, 2011). On the other hand, explanations offer customization possibilities, but at the cost of standardization (Berland and Reiser, 2009). For these reasons, we claim that, rather than 'one-shot' messages that users can only 'take or leave', similar to human-human

interaction explanations should be offered in the form of open and interactive dialogues, where users can question an explainer's account to expose possible inconsistency (Dunne et al., 2005) and mistakes (Lamche et al., 2014).

Additionally, we emphasized the connections between users' trust and their understanding of the causes of artificial agents' behavior. The possibility to question explanations and, in principle, the explainee's understanding allows users to gather deeper insights on artificial agents' actions maximize users' understanding, particularly if first explanatory attempts are not successful.

3.1 Interactive explanations and questionability

Some strategies exist to make explanations interactive and questionable. In principle, these can be applied both during or at the end of an explanation. For instance, Pieters proposes to organize artificial agents' explanations according to 'goals' and 'subgoals' (Pieters, 2011). If, for instance, the main goal of an explanation is to justify a specific decision, then a subgoal may be what Pieters calls 'transparency', that is gathering further information on how the explanation was constructed to make sure the agent didn't make errors (Pieters, 2011). Similarly, Madumal et al. developed an explanatory model that includes 'nested argumentation' modules (Madumal et al., 2018, 2019). These are dialogues 'nested' within an explanation that users can entertain with artificial agents. Importantly, such dialogues need not be related to the original question (Madumal et al., 2018, 2019).

'Examination phases' at the end of an explanation are yet another possibility (Dunne et al., 2005; Walton, 2011). Compared to other strategies, the main difference is that, in principle, an examination phase give both parties involved the chance to question and be questioned. The explainer's account can be questioned to evaluate if an explanation that sounds plausible is also truthful. Conversely, considering that people tend to overestimate their own 'knowledge retention' capacity (Keil, 2003; Pronin, 2009), the explainee's understanding may be tested as well. However, how exactly this should be done is an open question. In fact, finding the right balance between certainty of successful understanding and an overwhelming, inquisitorial number of questions is a challenging task (Papagni and Koeszegi, 2020; Walton, 2011). For this reason, some researchers propose to rely on the explainee's self-reporting (Madumal et al., 2019).

A reasonable compromise may be to ask the explainee to either present their own understanding of the explanation or pick the correct explanation from multiple choices. However, ultimately, whichever approach is the most suitable will depend on contextual affordances, such as how much time can be invested, or what are the consequences at stake.

While further empirical research is needed to validate this claim, early studies emphasize how interactivity and openness may improve explanations' quality and users' understanding. For instance, Alipour et al. (2020) conducted a study in the context of Visual Question Answering (VQA) to compare different explanations types in terms of users' predictions of the system's correctness.

Importantly, their results show that 'active attention explanations' (i.e., when the users can modify the system's original attention to generate different answers in the form of new attention maps) better supports users' confidence and trust towards the system, compared to other, more 'static' explanations.

3.2 Multi-modal explanations

In human-human interactions, explanations' content is mostly conveyed through natural language-based dialogues, typically in accordance with rules of cooperative conversation, such as the four 'Gricean maxims' (quality, quantity, relation and manner) (Grice, 1975; Hellström and Bensch, 2018; Hilton, 1990). Importantly, however, interactions with artificial agents offer complementary solutions.

Multi-modal explanations that use 'combined signals' (Engle, 1998) represent a promising direction and yet remain fairly uncharted terrain. Anjomshoae et al. identify six main modalities for artificial agents to convey explanations (Anjomshoae et al., 2019). In their analysis, text-based natural language explanations cover a significant part of the spectrum because, despite the availability of other means of communication, text encapsulates the richest (and perhaps clearest) semantic content. The other explanation modalities are, in order of importance: visualization, logs, expressive motions, expressive lights, and speech (Anjomshoae et al., 2019). While speech, which occupies the last position, is still based on natural language, what makes it less commonly used than other means is the difficulty of endowing an agent with it.

The availability of multiple channels does not necessarily imply that, to increase the chances of users understanding explanations, artificial agents should display all available information in the available formats at once. In fact, this 'infobesity' (Theodorou et al., 2016) might 'cognitively overload' users, who would then fail to understand (Lipton, 2016). Rather, the combination of different types of signals should be used to suit specific interaction contexts. For instance, Huk Park et al. (2018) conducted a study in the context of image classification graphic explanations of image recognition were accompanied by text-based captions describing fundamental parameters influencing the recognition process. The study's results indicate that the combination of visual and textual elements in the explanations enhanced the likelihood of users grasping the reasons behind specific predictions.

However, combined signals might not always be the most appropriate strategy. In certain cases, single-channel explanations may still be a better choice overall. For example, (Theodorou et al., 2016) consider the specific case of reactive planning and claim that, since artificial agents can take a great number of decisions per second, providing information verbally might be difficult for users to handle. Accordingly, they suggest that a graphical representation is a more efficient and direct way of making the information available even for less technical users, while preventing them from becoming overwhelmed (Theodorou et al., 2016). This again suggests that the choice of specific strategy to improve the quality of artificial agents' explanations strongly depends on the contextual

conditions within which interactions occur.

Multi-modality and interactivity represent two of the most promising strategies for ensuring a broad range of customization of explanations. Our final take on these strategies is that they do not need to be considered mutually exclusive alternatives. Instead, we claim that, depending on the contextual affordances, combining multi-modality and interactivity can offer even more reliable and personalized solutions to support users' understanding and trust development. To this extent, we close this section by showing how the combination of multi-modality and interactivity may work in two scenarios with significantly different interaction affordances.

3.3 Two cases for interactive, multi-modal explanations

The first example we present to demonstrate how interactivity and multi-modality can improve explanations discusses recommender systems in the context of online shopping. Recommender systems that suggest customers new products have become a very popular feature of shopping websites. Using techniques such as 'collaborative filtering', recommender systems provide customers with personalized suggestions about items to purchase. Filtering methods are usually based on implicit and explicit information about products or users similarities Leimstoll and Stormer (2007). This means that the more a customer interacts with the website by giving products rating (explicit information), clicking on specific objects or buying them (implicit information), the more accurate the recommendations become.

Implementing a combination of interactive and multi-modal explanations may contribute to users' perception of a personalized service. For instance, if a customer would want to know the reason for a book recommendation, an explainable recommender system may initially clarify that same book is similar to others that the customer has rated positively and that other readers with similar taste expressed positive opinions about it. However, the customer may ask further information before committing to spending money to buy the book. At this point, the system could provide additional details, for instance showing on a coarse level how the recommendation was generated, or displaying with graphic support how similar books 'scored' in terms of similarity with the customer's previous interactions. If a deeper level of insight would be requested by the customer, further information may be provided that show how each feature weighed in the process of generating the recommendation.

To extend our considerations on interactive, multi-modal explanations, the second case we discuss refers to using robots in search and rescue contexts. Replacing humans in 'dirty, dangerous and dull' jobs has historically been one of the main goals of robotics. To this extent, robots are meant to provide support with rescue missions in case of natural disasters like earthquakes Matsuno and Tadokoro (2004), or fires Wagoner et al. (2015) with tasks that include locating people trapped in buildings and guide them out safely, detect, avoid or extinguish fire.

The concerned people would likely be in a similar situation for the first time,

not knowing exactly what to do. Hence, it would be fundamental that the robot initially clarifies why it is there and how it may help (i.e., initial role explanation). As the robot guides people out, it may guide people towards the service staircase, rather than the main one. People could find this counterintuitive, for instance because the way to the main stairs is faster, and ask the robot why it is taking an alternative route. As timing would be an issue in such critical contexts, the robot would have to explain its decisions very quickly and effectively. Telling how its sensors detected high temperatures on the main staircase, or that rubble obstruct the stairs would likely be considered plausible explanations. If one would need further reassurance (reasonably so, given the high stakes), the robot might display a virtual map of the building showing the visualization of its sensor scans, or pictures of the rubble blocking the way. As chances of users correctly understanding the robot's explanations increase, the likeliness of users placing appropriate trust in the robot may also benefit, as well as human-robot collaboration in general.

While these scenarios only cover a minimal part of the possible applications of our approach explainability, the diversity of conditions that they represent outline the range of customization options enabled by contextual combinations of multi-modality and interactivity.

4 Conclusions

This paper discussed how explainability can support trust in human-agent interaction and from a time- and context-based perspective. To this extent, this paper focused on how to maximize the effect of explainability as a trust support strategy from the point of view of end-users, particularly non-expert ones, rather than from a technical stance. We first analyzed possible readings of trust relevant for this specific case. Specifically, the connections between trust, reliability and confidence were addressed. This perspective sought to emphasize the perception of risks and uncertainties implied in trust-based relationships, particularly before first interactions and after the occurrence of unexpected events. Furthermore, the study considered how the perceived role of 'third parties', such as the companies responsible for the development and distribution of artificial agents, can influence the trustworthiness of such agents.

Furthermore, we discussed how explanations may be generated and communicated to support (primarily) non-expert users' understanding of artificial agents' decision-making processes and trust towards them, with particular attention to those moments of an interaction in which trust is more at stake. Then, we graphically rendered our main findings into a model that displays the connections between trust, mental model construction and calibration and explanations throughout different phases of an interaction.

Thus, the main conclusions this paper draws are that artificial agents' trustworthiness is not a stable quality. As such, it can change as an interaction unfolds and can be influenced by several factors ranging from individual's disposition and artificial agents' capacity to perform according to their purpose,

to external factors such as other entities that may influence artificial agents' trustworthiness. Given that low levels of trust may hinder future interactions, making artificial agents explain their actions and decisions can effectively support trust over time, if explanations are properly tailored according to the users' needs and specific contextual affordances.

For future work, it is important to validate the main arguments of this paper in experimental studies. For instance, the effect of an artificial agent's explanations (or lack thereof) at the beginning of an interaction and after a mistake may be tested in terms of effect on the agent's trustworthiness and understandability. Likewise, different types of explanations may be tested in relation to different users' characteristics and contexts of interaction. Finally, how the proposed approach to explainability fit different techniques to generate explanation may be addressed by future work.

Paper title: May I explain? Explainability as a Trust Support Strategy for Artificial Agents

References

Adadi A, Berrada M (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6:52138–52160, DOI 10.1109/ACCESS.2018.2870052

Alipour K, Schulze JP, Yao Y, Ziskind A, Burachas G (2020) A study on multimodal and interactive explanations for visual question answering. arXiv preprint arXiv:200300431

Andras P, Esterle L, Guckert M, Han TA, Lewis PR, Milanovic K, Payne T, Perret C, Pitt J, Powers ST, Urquhart N, Wells S (2018) Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems. IEEE Technology and Society Magazine 37(4):76–83, DOI 10.1109/MTS.2018.2876107

Anjomshoae S, Najjar A, Calvaresi D, Främling K (2019) Explainable Agents and Robots: Results from a Systematic Literature Review. International Foundation for Autonomous Agents and Multiagent Systems, URL http://dl.acm.org/citation.cfm?id=3306127.3331806

Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, et al. (2018) Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:181001943

Berland LK, Reiser BJ (2009) Making sense of argumentation and explanation. Science education 93(1):26–55

Cawsey A (1993) User modelling in interactive explanations. User Modeling and User-Adapted Interaction 3(3):221–247

Chen Y, Wu F, Shuai W, Wang N, Chen R, Chen X (2015) Kejia robot–an attractive shopping mall guider. In: International Conference on Social Robotics, Springer, pp 145–154

Coeckelbergh M (2018) How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of icts in terms of magic and performance and taking a relational and narrative turn. Ethics and Information Technology 20(2):71–85

De Graaf MM, Malle BF (2017) How people explain action (and autonomous intelligent systems should too). In: 2017 AAAI Fall Symposium Series

De Visser EJ, Peeters MM, Jung MF, Kohn S, Shaw TH, Pak R, Neerincx MA (2020) Towards a theory of longitudinal trust calibration in human–robot teams. International journal of social robotics 12(2):459–478

Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:170208608

Dunne PE, Doutre S, Bench-Capon T (2005) Discovering inconsistency through examination dialogues. In: Proceedings of the 19th international joint conference on Artificial intelligence, pp 1680–1681

Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP (2003) The role of trust in automation reliance. International journal of human-computer studies 58(6):697–718

Elangovan A, Auer-Rizzi W, Szabo E (2007) Why don't i trust you now? an attributional approach to erosion of trust. Journal of Managerial Psychology

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

Engle RA (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, pp 321–326

Ferreira JJ, Monteiro MdS (2020) Do ml experts discuss explainability for ai systems? a discussion case in the industry for a domain-specific solution. arXiv preprint arXiv:200212450

Fossa F (2019) " i don't trust you, you faker!" on trust, reliance, and artificial agency

Fulmer CA, Gelfand MJ (2012) At what level (and in whom) we trust: Trust across multiple organizational levels. Journal of management 38(4):1167–1230

Gefen D (2000) E-commerce: the role of familiarity and trust. Omega 28(6):725–737

Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a "right to explanation". AI magazine 38(3):50–57

de Graaf MM, Malle BF, Dragan A, Ziemke T (2018) Explainable robotic systems. In: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp 387–388

Grice HP (1975) Logic and conversation. In: Speech acts, Brill, pp 41–58

Gunning D (2017) Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web 2(2)

Gunning D, Aha DW (2019) Darpa's explainable artificial intelligence program. AI Magazine 40(2):44–58

Hagras H (2018) Toward Human-Understandable, Explainable AI. Computer 51(9):28–36, DOI 10.1109/MC.2018.3620965

Hancock PA, Billings DR, Schaefer KE, Chen JY, De Visser EJ, Parasuraman R (2011) A meta-analysis of factors affecting trust in human-robot interaction. Human factors 53(5):517–527

Harman GH (1965) The inference to the best explanation. The philosophical review 74(1):88–95

Hellström T, Bensch S (2018) Understandable robots - What, Why, and How. Paladyn, Journal of Behavioral Robotics 9(1):110–123, DOI 10.1515/pjbr-2018-0009

Hesslow G (1988) The problem of causal selection. Contemporary science and natural explanation: Commonsense conceptions of causality pp 11–32

Hilton DJ (1990) Conversational processes and causal explanation. Psychological Bulletin 107(1):65

Hilton DJ, Slugoski BR (1986) Knowledge-based causal attribution: The abnormal conditions focus model. Psychological review 93(1):75

Hilton DJ, McClure J, Sutton RM (2010) Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? European Journal of Social Psychology 40(3):383–400

Hleg A (2019) Ethics guidelines for trustworthy ai. B-1049 Brussels

Ho N, Sadler GG, Hoffmann LC, Zemlicka K, Lyons J, Fergueson W, Richardson C, Cacanindin A, Cals S, Wilkins M (2017) A longitudinal field study of auto-gcas acceptance and trust: First-year results and implications. Journal of Cognitive Engineering and Decision Making 11(3):239–251

Hois J, Theofanou-Fuelbier D, Junk AJ (2019) How to Achieve Explainability and Transparency in Human AI Interaction. In: Stephanidis C (ed) HCI International 2019 - Posters, vol 1033, Springer International Publishing, Cham, pp 177–183, DOI 10.1007/978-3-030-23528-4$_2$5

Holliday D, Wilson S, Stumpf S (2016) User trust in intelligent systems: A journey over time. In: Proceedings of the 21st international conference on intelligent user interfaces, pp 164–168

Huk Park D, Anne Hendricks L, Akata Z, Rohrbach A, Schiele B, Darrell T, Rohrbach M (2018) Multimodal explanations: Justifying decisions and pointing to the evidence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 8779–8788

Im I, Hong S, Kang MS (2011) An international comparison of technology adoption: Testing the utaut model. Information & management 48(1):1–8

Jacovi A, Marasović A, Miller T, Goldberg Y (2021) Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp 624–635

Kahneman D, Tversky A (1981) The simulation heuristic. Tech. rep., Stanford Univ CA Dept of Psychology

Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J (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, pp 1–14

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

Kerschner C, Ehlers MH (2016) A framework of attitudes towards technology in theory and practice. Ecological Economics 126:139–151

Kulesza T, Stumpf S, Burnett M, Yang S, Kwan I, Wong WK (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, pp 3–10

Lakkaraju H, Bastani O (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, pp 79–85

Lamche B, Adıgüzel U, Wörndl W (2014) Interactive explanations in mobile shopping recommender systems. In: Joint Workshop on Interfaces and Human Decision Making in Recommender Systems, vol 14

Lankton NK, McKnight DH, Tripp J (2015) Technology, humanness, and trust: Rethinking trust in technology. Journal of the Association for Information Systems 16(10):1

Lee JD, See KA (2004) Trust in automation: Designing for appropriate reliance. Human factors 46(1):50–80

Leimstoll U, Stormer H (2007) Collaborative recommender systems for online shops. In: Proceedings of the 2007 Americas Conference on Information Systems (AMCIS), vol 156

Lewicki RJ, Brinsfield C (2017) Trust repair. Annual Review of Organizational Psychology and Organizational Behavior 4:287–313

Lewis D (1986) Causal explanation

Li X, Hess TJ, Valacich JS (2008) Why do we trust new technology? a study of initial trust formation with organizational information systems. The Journal of Strategic Information Systems 17(1):39–71

Lipton ZC (2016) The Mythos of Model Interpretability. arXiv:160603490 [cs, stat] 1606.03490

Lipton ZC, Steinhardt J (2018) Troubling Trends in Machine Learning Scholarship. arXiv URL https://arxiv.org/abs/1807.03341, 1807.03341

Lockey S, Gillespie N, Holm D, Someh IA (2021) A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. In: Proceedings of the 54th Hawaii International Conference on System Sciences, p 5463

Lomas M, Chevalier R, Cross EV, Garrett RC, Hoare J, Kopack M (2012) Explaining robot actions. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, pp 187–188

Lombrozo T (2006) The structure and function of explanations. Trends in Cognitive Sciences 10(10):464–470, DOI 10.1016/j.tics.2006.08.004

Lombrozo T (2007) Simplicity and probability in causal explanation. Cognitive psychology 55(3):232–257

Luhmann N (2000) Familiarity, confidence, trust: Problems and alternatives. Trust: Making and breaking cooperative relations 6(1):94–107

Luhmann N (2018) Trust and power. John Wiley & Sons

Lyon F, Möllering G, Saunders MN (2015) Introduction. researching trust: the ongoing challenge of matching objectives and methods. In: Handbook of research methods on trust, Edward Elgar Publishing

Madumal P, Miller T, Vetere F, Sonenberg L (2018) Towards a grounded dialog model for explainable artificial intelligence. arXiv preprint arXiv:180608055

Madumal P, Miller T, Sonenberg L, Vetere F (2019) A grounded interaction protocol for explainable artificial intelligence. arXiv preprint arXiv:190302409

Malle BF (2011) Attribution theories: How people make sense of behavior. Theories in social psychology 23:72–95

Malle BF, Knobe JM, Nelson SE (2007) Actor-observer asymmetries in explanations of behavior: New answers to an old question. Journal of personality and social psychology 93(4):491

van Maris A, Lehmann H, Natale L, Grzyb B (2017) The influence of a robot's embodiment on trust: A longitudinal study. In: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp 313–314

Matsuno F, Tadokoro S (2004) Rescue robots and systems in japan. In: 2004 IEEE international conference on robotics and biomimetics, IEEE, pp 12–20

Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267:1–38, DOI 10.1016/j.artint.2018.07.007

Mohseni S, Zarei N, Ragan ED (2018) A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. arXiv:181111839 [cs] 1811.11839

Mollering G (2006) Trust: Reason, routine, reflexivity. Emerald Group Publishing

Molnar C (2020) Interpretable Machine Learning. Lulu. com

Morris MG, Venkatesh V (2000) Age differences in technology adoption decisions: Implications for a changing work force. Personnel psychology 53(2):375–403

Niemelä M, Heikkilä P, Lammi H (2017) A social service robot in a shopping mall: expectations of the management, retailers and consumers. In: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp 227–228

O'Leary DE (2019) Google's duplex: Pretending to be human. Intelligent Systems in Accounting, Finance and Management 26(1):46–53

O'neill O (2002) Autonomy and trust in bioethics. Cambridge University Press

Papagni G, Koeszegi S (2020) Understandable and trustworthy explainable robots: A sensemaking perspective. Paladyn, Journal of Behavioral Robotics 12(1):13–30

Papagni G, Koeszegi S (2021) A pragmatic approach to the intentional stance semantic, empirical and ethical considerations for the design of artificial agents. Minds and Machines 31(4):505–534

Peirce CS (1997) Pragmatism as a principle and method of right thinking: The 1903 Harvard lectures on pragmatism. SUNY Press

Pieters W (2011) Explanation and trust: what to tell the user in security and ai? Ethics and information technology 13(1):53–64

Preece A, Harborne D, Braines D, Tomsett R, Chakraborty S (2018) Stakeholders in explainable ai. arXiv preprint arXiv:181000184

Pronin E (2009) The introspection illusion. Advances in experimental social psychology 41:1–67

Pu P, Chen L (2007) Trust-inspiring explanation interfaces for recommender systems. Knowledge-Based Systems 20(6):542–556

Quinn DB, Pak R, de Visser EJ (2017) Testing the efficacy of human-human trust repair strategies with machines. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications Sage CA: Los Angeles, CA, vol 61, pp 1794–1798

Riedl MO (2019) Human-centered artificial intelligence and machine learning. Human Behavior and Emerging Technologies 1(1):33–36

Robinette P, Li W, Allen R, Howard AM, Wagner AR (2016) Overtrust of robots in emergency evacuation scenarios. In: 2016 11th ACM/IEEE international conference on human-robot interaction (HRI), IEEE, pp 101–108

Robinette P, Howard AM, Wagner AR (2017) Effect of robot performance on human–robot trust in time-critical situations. IEEE Transactions on Human-Machine Systems 47(4):425–436

Rossi A, Dautenhahn K, Koay KL, Walters ML, Holthaus P (2020) Evaluating people's perceptions of trust in a robot in a repeated interactions study. In: International Conference on Social Robotics, Springer, pp 453–465

Rotter JB (1971) Generalized expectancies for interpersonal trust. American psychologist 26(5):443

Rudin C (2018) Please Stop Explaining Black Box Models for High Stakes Decisions. arXiv URL https://arxiv.org/abs/1811.10154, 1811.10154

Salem M, Lakatos G, Amirabdollahian F, Dautenhahn K (2015) Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust. In: 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, pp 1–8

Schoorman FD, Mayer RC, Davis JH (2007) An integrative model of organizational trust: Past, present, and future. Academy of Management Review 32(2):344–354

Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. International Journal of Human-Computer Studies 146:102551

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

Simpson JA (2007) Foundations of interpersonal trust. Social psychology: Handbook of basic principles 2:587–607

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

Taddeo M, Floridi L (2011) The case for e-trust. Ethics and Information Technology 13(1):1–3

Thagard P (1989) Explanatory coherence. Behavioral and brain sciences 12(3):435–502

Theodorou A, Wortham RH, Bryson JJ (2016) Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots. the University of Bath's research portal

Venkatesh V, Morris MG, Ackerman PL (2000) A longitudinal field investigation of gender differences in individual technology adoption decision-making processes. Organizational behavior and human decision processes 83(1):33–60

Wagner AR, Borenstein J, Howard A (2018) Overtrust in the robotic age. Communications of the ACM 61(9):22–24

Wagoner A, Jagadish A, Matson ET, EunSeop L, Nah Y, Tae KK, Lee DH, Joeng JE (2015) Humanoid robots rescuing humans and extinguishing fires for cooperative fire security system using harms. In: 2015 6th International Conference on Automation, Robotics and Applications (ICARA), IEEE, pp 411–415

Walton D (2011) A dialogue system specification for explanation. Synthese 182(3):349–374

Wang T (2019) Gaining free or low-cost interpretability with interpretable partial substitute. In: International Conference on Machine Learning, PMLR, pp 6505–6514

Wang T, Lin Q (2021) Hybrid predictive models: When an interpretable model collaborates with a black-box model. Journal of Machine Learning Research 22(137):1–38

Weick KE, Sutcliffe KM, Obstfeld D (2005) Organizing and the process of sensemaking. Organization science 16(4):409–421

Wiegand G, Schmidmaier M, Weber T, Liu Y, Hussmann H (2019) I drive-you trust: Explaining driving behavior of autonomous cars. In: Extended abstracts of the 2019 chi conference on human factors in computing systems, pp 1–6

Wilkenfeld DA, Lombrozo T (2015) Inference to the best explanation (ibe) versus explaining for the best inference (ebi). Science & Education 24(9-10):1059–1077

Woodcock C, Mittelstadt B, Busbridge D, Blank G, et al. (2021) The impact of explanations on layperson trust in artificial intelligence–driven symptom checker apps: Experimental study. Journal of medical Internet research 23(11):e29386

Zafari S, Koeszegi ST (2018) Machine agency in socio-technical systems: A typology of autonomous artificial agents. In: 2018 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), IEEE, pp 125–130

Zaheer A, McEvily B, Perrone V (1998) Does trust matter? exploring the effects of interorganizational and interpersonal trust on performance. Organization science 9(2):141–159

Zemla JC, Sloman S, Bechlivanidis C, Lagnado DA (2017) Evaluating everyday explanations. Psychonomic bulletin & review 24(5):1488–1500

Zou J, Schiebinger L (2018) Ai can be sexist and racist—it's time to make it fair

Zucker LG (1987) Institutional theories of organization. Annual review of sociology 13(1):443–464

CHAPTER 4

Paper3

At the time of the submission of this dissertation, the following paper is still under review at the "*ACM Transactions on Recommender Systems*" journal.

Explanations for trust development and trust restoration for assistive recommender systems: a longitudinal study

SETAREH ZAFARI*, Austrian Institute of Technology, Austria
JESSE DE PAGTER* and GUGLIELMO PAPAGNI*, Institute of Management Science, TU Wien, Austria
ALISCHA ROSENSTEIN, UX Concepts (RD/ECC), Mercedes-Benz AG, Germany
MICHAEL FILZMOSER, Institute of Management Science, TU Wien, Austria
SABINE T. KOESZEGI, Institute of Management Science, TU Wien, Austria

This article reports on a longitudinal experiment in which the influence of an assistive system's malfunctioning and transparency on trust was examined over a period of seven days. To this end, we simulated the system's personalized recommender features to support participants with the task of learning new texts and taking quizzes. Using a 2×2 mixed design, the system's malfunctioning (correct vs. faulty) and transparency (with vs. without explanation) were manipulated as between-subjects variables, whereas exposure time was used as a repeated-measures variable. A combined qualitative and quantitative methodological approach was used to analyze the data from 171 participants. Our results show that participants perceived the system making a faulty recommendation as a trust violation. Additionally, system transparency (explanation of malfunction) led to faster trust restoration. Whereas participants did not always access or use explanations of malfunctions, qualitative analyses indicated that the mere availability of explanations can add to the experience of trustworthiness.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **User studies**; • **Applied computing** → *Computer-assisted instruction.*

Additional Key Words and Phrases: Explainability, Transparency, Trust development, Trust restoration, System malfunction

ACM Reference Format:
Setareh Zafari, Jesse de Pagter, Guglielmo Papagni, Alischa Rosenstein, Michael Filzmoser, and Sabine T. Koeszegi. 2023. Explanations for trust development and trust restoration for assistive recommender systems: a longitudinal study. *ACM Trans. Recomm. Syst.* XXX, XXX, Article XXX (XXX 2023), 26 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Trust is a fundamental concept in human relationships, as people's behavior depends on their trust in each other. Hence, trust is investigated from a wide range of perspectives. Important

*All three authors contributed equally to this research.

Authors' addresses: Setareh Zafari, setareh.zafari@ait.ac.at, Austrian Institute of Technology, Giefinggasse 2, Vienna, Austria; Jesse de Pagter, jesse.de.pagter@tuwien.ac.at; Guglielmo Papagni, guglielmo.papagni@tuwien.ac.at, Institute of Management Science, TU Wien, Theresianumgasse 27, Vienna, Austria, 1040; Alischa Rosenstein, UX Concepts (RD/ECC), Mercedes-Benz AG, Leibnizstrasse 2, Stuttgart, Germany, 71032; Michael Filzmoser, Institute of Management Science, TU Wien, Theresianumgasse 27, Vienna, Austria, 1040; Sabine T. Koeszegi, Institute of Management Science, TU Wien, Theresianumgasse 27, Vienna, Austria, 1040.

2770-6699/2023/XXX-ARTXXX $15.00
https://doi.org/XXXXXXX.XXXXXXX

examples include antecedents of trust [21], cognitive and emotional components of trust [41], trust in organizations [57, 68], and trust in interpersonal relationships [52, 59].

In recent decades, artificial intelligence (AI) has been increasingly used in a growing number of application contexts. As many applications of AI-based technologies, ranging from email services to online banking, social media and recommender systems, affect people's everyday lives in a direct or indirect way, the concept of trust has come to occupy a central position in academic and institutional discussions related to AI-based autonomous systems. Specifically for recommender systems as the "frontiers of Human-centered AI research" [20, p.1], theorists and researchers aim to understand the dynamics of trust formation with respect to these systems, how trust changes over time [33, 55, 58], and whether and how these processes relate to trust in human-human interaction.

Transparency,or the lack thereof, is considered one of the idiosyncratic yet most relevant features of AI that may influence how much trust people place in it [5, 28, 56]. This is particularly the case with complex models such as neural networks (black-box models), which are currently increasing in popularity [1]. In this regard, researchers argue that making the causal chains behind models' decisions interpretable and transparent is likely to help people understand the rationales behind those decisions and calibrate their expectations and trust [10, 23, 47]. This, in turn, increases the chances that users decide to interact further with a system [49]. Given that trust is dynamic and changes throughout different phases of interactions, still-open questions concern how exactly trust forms and evolves in the context of repeated interactions with AI-based systems, as well as the conditions under which transparency affects trust in these systems. The experimental work presented here contributes to the literature by investigating trust dynamics in the context of repeated interaction with an assistive system. To this end, we simulated the system's personalized recommendations in order to assist users preparing for quizzes by providing them with recommendations on which portions of text to focus on. Specifically, this study focuses on comparing participants' trust ratings at the beginning, over the course of the study, and after a system malfunction. Furthermore, the system provided explanations of how it functions to one group of participants, while another group was not provided with this explanation. Trust ratings between these conditions are also compared. Finally, trust in the system at the very end of the study is also measured and compared. Our study shows that, even after the system has proven its reliability, a faulty recommendation is perceived as a trust violation. Accordingly, participants who experienced the system's malfunction attributed significantly lower trust to it than those who interacted with an always accurate system. Furthermore, in the case of a faulty recommendation, providing explanations to clarify the causes of the system's malfunction results in significantly faster trust recovery compared to when no explanation is offered by the system.

The remainder of this paper is divided into five parts. Section 2 discusses previous work related to the notion of trust as a dynamic process, connecting it to the concept of explainability, while identifying open challenges and questions. Then, Section 3 describes the methodology and design of the 2x2 study in which the system's accuracy and explainability were manipulated to investigate how trust in the system is affected. Section 4 presents the results from the quantitative and qualitative analysis. The study's contributions to the literature on trust and explainability are then discussed in Section 5, together with final considerations and limitations in section 6.

2 THEORETICAL BACKGROUND AND HYPOTHESES

Definitions of trust have repeatedly emphasized certain elements. Namely, trust implies a trustor who is willing to be vulnerable, face risks and uncertainties in expectation that the trustee will provide support in achieving specific goals [19, 59]. As such, trust is a fundamental phenomenon that characterizes human relationships on multiple levels. Trust in technology represents just one of these levels (albeit a multifaceted one), and researchers emphasize how trust plays a role in

determining technology acceptance [37, 58]. In this respect, trust towards an assistive system can be operationalized as the probability of an individual following the system's recommendations, predictions, and classifications [56].

2.1 Initial trust

The dynamic nature of trust necessitates studying it at different moments of an interaction [11, 39, 42, 46]. Certain factors may influence people's initial trust in new technologies before any interaction takes place. As antecedents of trust, individuals' characteristics, environmental factors and features of the technology in question play a role in determining people's initial trust towards new technologies [58]. Taken together, these factors contribute to determining people's initial attitude and expectations, or else initial trust would be a 'blind leap of faith' [36].

Environmental factors include social and cultural background and institutional cues. The latter is particularly relevant for AI-based technologies, as it refers to entities that are involved in the introduction of new technologies, such as developers and experts, companies that market the technology, and national and international organizations that contribute to shaping the narratives around new technologies [4, 35, 42]. Before any interaction is established, institutional cues can determine whether people perceive new technologies as benevolent or malicious [32, 60]. To this extent, [44] show how AI experts' expressions of pessimistic positions on AI on Twitter, perhaps more significantly than the technology's actual progress, influenced people's perception of growing risks, both existential and not (e.g., job replacement), related to the recent surge of AI applications.

Human factors refer to the disposition to trust, propensity to take risks, individual abilities and personality traits [29, 55, 58]. [8] investigated the effect of cultural differences and personality traits on trust in automation and demonstrated that both play a role, individually and combined. Interestingly, with respect to the five-factor model of personality, their results highlight how high agreeableness and conscientiousness correspond with high levels of trust in automation. Since most definitions of trust include risk taking as a core aspect, some existing studies focus specifically on how the trust and risk dimensions are intertwined. For instance, [2] report on a study conducted to evaluate the effects of risk perception on trust in autonomous vehicles. They found that not only did interacting with an autonomous vehicle in a risky scenario significantly reduce participants' trust and delegation of control to the vehicle, but also that initial trust was significantly higher than trust levels after interacting with the vehicle in high-risk conditions. Another study investigating risk aversion in accommodation context suggests that trust and risk are constructs that may be closely related in personal exchange contexts [18]. Accordingly, the authors found that risk-averse individuals are more likely to weigh the loss associated with trusting a system over the potential gain. Thus, we expect risk attitudes plays an important role in trust development.

Concerning technological features, [33] identify three factors that can influence people's trust in automation. These are performance, process and purpose. Researchers argue that AI-based system may be considered trustworthy if it acts within the 'contractual preconditions' of its use [28], that is, if an AI-based system successfully performs in accordance with its purposes, which are contextually recognized by users.

However, before or at the beginning of an interaction, it is difficult for people to judge whether an AI-based system will perform in accordance with its purposes. This means that initial trust is likely not based on the AI-based system's actual capabilities. Rather, it is influenced by individuals' background and disposition and how the technology is presented by external entities. This may translate into unreasonably high or low levels of initial trust [15, 30]. In this regard, explainability may help to calibrate initial trust by compensating for the lack of previous experience necessary to evaluate performance in accordance with purposes [4, 36, 58]. As [4] note, during the first phases

of the adoption of specific AI-based technologies, explaining how they operate may reduce users' perception of risks.

Accordingly, we propose:

H1a: Transparency by means of explanations about the system's inner workings leads to higher initial trust levels.

H1b: Participants with higher risk propensity will tend to have higher initial trust towards the system.

2.2 Trust development over time

Once initial trust is established and the interaction with an AI-based system proceeds, people are unlikely to completely lose trust without a specific reason. However, researchers suggest that trust dynamics change gradually [58] and that initial trust levels usually adjust after an interaction begins as the result of a calibration of individuals' attitudes and other factors involved in determining initial trust, intertwined with an AI-based system's behavior [24, 26, 27, 39]. Recalling Lee and See's model, as an interaction unfolds, an AI-based system will likely be considered trustworthy if it performs in accordance with its purpose and the 'contracts' established with users [28, 33]. For an AI-based system to be considered reliable, behavioral consistency over time is required. In fact, reliability is a property that can be attributed to a system only in relation to its past performance [17, 45]. In turn, when an AI-based system proves reliable, people grow confident in its capacity and trust and familiarity stabilize [31, 38, 66].

Researchers suggest that in this phase, explanations may be superfluous [9, 13], if not detrimental for trust [56]. For instance, a study conducted to test how various explanation types satisfy explanation quality criteria showed how counterfactual explanations, a type of explanation very close to human experience (see [43]), did not improve trust calibration among participants [65]. Moreover, [6] argue that explanations may reveal an AI-based system's limited capabilities, breaking the illusion of intelligence and hindering trust. On the other hand, too complex explanations (i.e., not calibrated to users' expertise) have been shown to undermine acceptance of recommender systems [25].

However, other studies point out that explanations during an interaction may help people make sense of specific decisions or predictions generated by AI-based systems [50] and are therefore fundamental to continuous trust calibration [58]. For instance, Jacovi et al. claim that the benefits of making AI explainable include increasing an AI-based system's trustworthiness, the trust people place in a trustworthy system, as well as the distrust triggered by a non-trustworthy system (i.e., trust calibration) [28]. Turning to empirical studies addressing continuous trust and explanations, [34] report on an experiment in which an explainable AI was used to support decision-making in a high-stakes context. Participants were introduced to an app-based system for recognizing specific mushrooms as edible or not. Their results indicate that explanations of how the AI-based system worked did not improve task performance, but explanations of specific predictions did. Interestingly, results from another study on explanations by autonomous vehicles show how explanations' timing plays a central role in determining users' trust. The researchers found that explanations provided before the vehicle acted had a positive influence on participants' attribution of trust, while explanations that were given after a specific action did not affect trust ratings [14].

Accordingly, we propose:

H2: System transparency by means of explanations provided throughout the study accelerates trust development as compared to non-transparent systems.

2.3 Trust violation and restoration

As part of the dynamic nature of trust, it may be that, during an interaction with an AI-based system, after the system proves reliable, something happens that compromises people's trust in it. Not only that, if people lose trust in an AI-based system due to a specific event, acceptance of the system and future interactions with it may also be hindered. In their taxonomy of events that can cause trust breaches, [62] identify four types of failures related to poor design choices, system failure, behavior that goes against users' expectations, and users' misbehavior.

Several studies support the idea that different types of failure may affect trust in different ways. For instance, in an empirical study of real-time evaluations of trust, Desai et al. found that mistakes by a robot that occur early in an interaction (i.e., before the robot has fully established its reliability) have more negative effects on trust than mistakes that occur later [12]. In another experiment [51], participants could choose whether to follow a robot to escape a dangerous situation. The results indicate how the robot's failures triggered significantly lower trust ratings compared to the condition in which the robot carried out the task successfully. Robotic failures also reduced motivation to interact further with the robot, which the authors suggest may indicate an increased perception of risk [51]. Another study that investigated the effect of a robot's performance (i.e., successful or faulty) on participants' trust produced similar results in that the faulty robot was considered significantly less trustworthy and reliable than the one that performed successfully. However, in this case, the robot's mistakes did not affect participants' willingness to follow the robot's instructions [53]. Results from other studies suggest that people's perception of the risk involved in the interaction as well as their individual disposition towards takings risks may explain this discrepancy [18, 48].

In their taxonomy, [62] also suggest trust restoration strategies. They identify approaches such as apologies, promises, remedial trustworthy behavior, and explanations that have the power to repair trust after a violation. Unlike many of these trust restoration strategies, explainability comes with two major advantages. First, as previously noted, explanations may support trust calibration not only in the case of a system's mistakes or malfunctioning, but also at the beginning of and throughout an interaction. Second, if properly tailored, explanations can shed light on the causes of a mistake, rather than just offering a restructuring of the relationship [46]. Furthermore, studies suggest that transparency achieved by means of explanations may not only restore trust after a violation, but also dampen it in case people over-trust a non-trustworthy system [11, 28].

Several empirical studies corroborate the idea that explanations, particularly if provided after mistakes or malfunctions, can mitigate negative effects on trust. For instance, [15] conducted a series of studies to investigate reliance on and trust in an automated decision aid by manipulating the automated aid's accuracy, task completion feedback, transparency about potential mistakes and other features. In one of these studies, they found that when the system provided explanations for why mistakes (i.e., false alarms or missed targets) might have occurred, trust in the system and reliance on its decisions increased in comparison to when the system provided no explanations. However, they conclude that high levels of trust in a faulty system may be dangerous, even if the system can explain its mistakes. Accordingly, they suggest that more informative explanations (e.g., about how the system takes correct decisions) and instructions about how the system operates may mitigate such unwanted effects.

In another study, [63] tested the effect of different explanation types (confidence-level explanation, observation explanation and no explanations) on trust in a simulated human-robot interaction study in the context of reconnaissance missions. Importantly, they also manipulated the robot's ability level (low and high ability). Their results show that both types of explanations yielded higher trust ratings compared to the no explanation condition, particularly in the low-ability case. Their

interpretation is that when interacting with a robot that is not always accurate, people have to pay more attention to explanations, which in turn become fundamental for people to decide whether to trust the robot or not. Alternatively, if a robot's performance is always accurate and the robot is hence perceived as reliable, explanations are not likely to positively affect trust.

Accordingly, we propose:

H3a: A system malfunction leads to a significant trust reduction.

H3b: System transparency by means of an explanation accelerates trust restoration after a malfunction.

H3c: Trust restoration depends on people's attitude towards risk.

3 METHOD

3.1 Experimental design

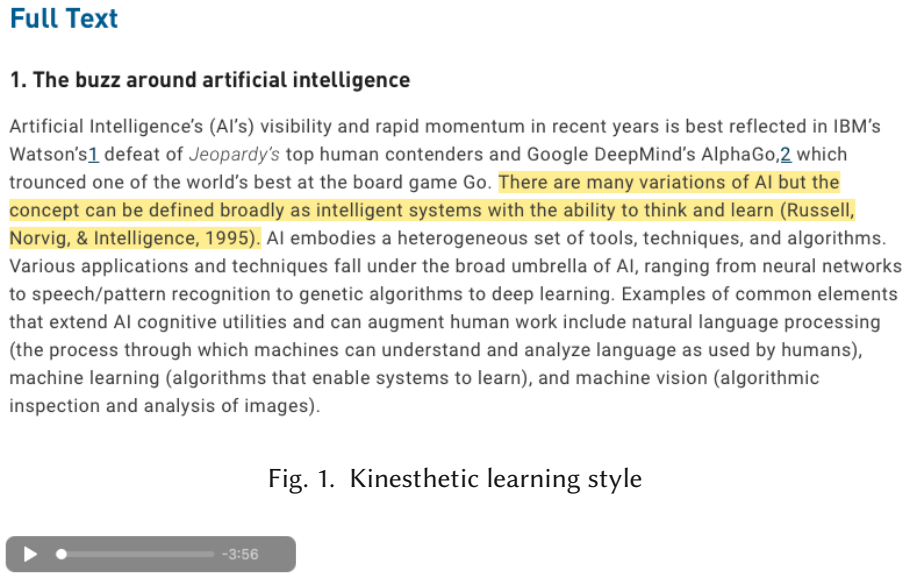
To test our hypotheses, a 2×2 mixed design with the following independent variables was implemented: system malfunctioning (correct/faulty), transparency (with/without explanation) and exposure time (measured over seven days). The system's malfunction and transparency were manipulated as between-subjects variables and exposure time was a within-subjects variable. For this purpose, we mimicked an abstract-generating recommender system named PLANT.

3.1.1 Use case: PLANT as a personalized recommender system. The system's main goal in our study was to support participants with the task of learning new technology-related texts and taking multiple short quizzes (five questions each). To meet its goal, the system provided personalized recommendations on the most relevant parts of a long text (i.e. abstracting support) to prepare participants for upcoming quizzes about the full text. While participants always had the option to access the full texts, accepting PLANT's recommendations allowed them to save time in preparing for the quizzes.

Importantly, the system's functionalities were mimicked through Wizard of Oz methodology. When participants were introduced to the system, they were explained that the recommendations were generated through the system's Natural Language Processing (NLP) algorithms, while in reality the researchers behind the project produced and controlled them. Furthermore, participants were told that the goal of the experiment was to test the system's functionalities in order to provide feedback to the developers. The mimicked nature of PLANT ensured the controllability of the study, as no actual malfunctions could occur.

Given that PLANT was a personalized recommender system with an abstracting function, one of its key features was the range of customization options. These included alerts and notifications, via either email or (optionally) SMS, with reminders of upcoming quizzes and suggestions to change the scheduling and timing of one's preparation. Additionally, participants could receive performance-based insights into their use of the recommendations, switch between 'light' and 'dark' themes for the interface and personalize the text font. Perhaps more importantly, participants could personalize their learning style by choosing among four different options (see Figures 1, 2, 3, 4). Specifically, these were:

- Kinesthetic: Full text with highlights.
- Auditory: Reading and listening to the summary.
- Reading/Writing: Bullet points.
- Visual: Graphical representation.



Full Text

1. The buzz around artificial intelligence

Artificial Intelligence's (AI's) visibility and rapid momentum in recent years is best reflected in IBM's Watson's1 defeat of *Jeopardy's* top human contenders and Google DeepMind's AlphaGo,2 which trounced one of the world's best at the board game Go. There are many variations of AI but the concept can be defined broadly as intelligent systems with the ability to think and learn (Russell, Norvig, & Intelligence, 1995). AI embodies a heterogeneous set of tools, techniques, and algorithms. Various applications and techniques fall under the broad umbrella of AI, ranging from neural networks to speech/pattern recognition to genetic algorithms to deep learning. Examples of common elements that extend AI cognitive utilities and can augment human work include natural language processing (the process through which machines can understand and analyze language as used by humans), machine learning (algorithms that enable systems to learn), and machine vision (algorithmic inspection and analysis of images).

Fig. 1. Kinesthetic learning style

Summary

1. The buzz around artificial intelligence

There are many variations of AI but the concept can be defined broadly as intelligent systems with the ability to think and learn.

1.1. How we talk about AI

Whereas the recent hyperbole surrounding AI and other cognitive technologies has led many to believe that machines will soon outthink humans and replace them in the workplace, others see the concern around AI as another overhyped proposition.

Fig. 2. Auditory learning style

Bullet points

1. The buzz around artificial intelligence

- There are many variations of AI but the concept can be defined broadly as intelligent systems with the ability to think and learn.

1.1. How we talk about AI

- Whereas the recent hyperbole surrounding AI and other cognitive technologies has led many to believe that machines will soon outthink humans and replace them in the workplace, others see the concern around AI as another overhyped proposition.

Fig. 3. Reading/writing learning style

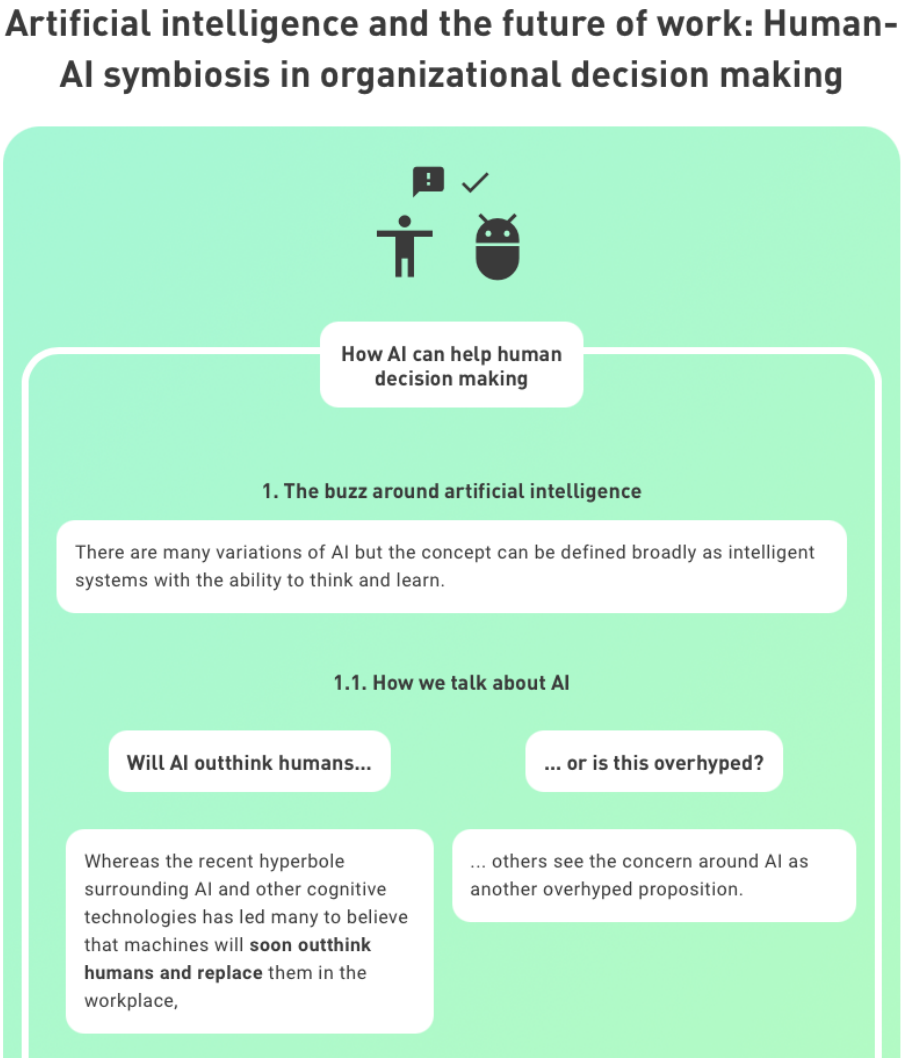


Fig. 4. Visual learning style

What each of the four learning styles respectively entails will now be briefly described. The 'kinesthetic' learning style presented the full text with essential sentences highlighted. The 'auditory' style was a shortened version of the full text consisting of the highlighted text passages only. Additionally, the summary featured headings and sub-headings corresponding to the sections of the full text. Participants could also listen to an auditory version of the summarized text. The 'reading/writing' style was adapted from the summary, with the essential passages slightly changed or shortened to create a list of appropriate bullet points. The headings and sub-headings of the summary were also shown in the list of bullet points. The contents of the 'visual' were derived from the bullet points by shortening and modifying the essential passages. In order to distinguish between different parts of the text, the sections were colored differently and icons were used to support the text. The infographic would automatically switch to a different set of colors if the user turned on the dark theme. The highlights, summary and bullet points were designed by the researchers responsible for the study using the edit tab in the back-end of the web application. The visualization was created with CSS classes based on the Flexbox Grid [1] system.

[1] http://flexboxgrid.com, last accessed September 2021

Based on how participants answered an initial questionnaire, the VARK Questionnaire Version 8.01 [2], PLANT suggested one of these learning styles to each user. However, participants did not have to follow the system's recommendation and it was up to them to decide at this point which learning style they felt most comfortable proceeding with. Importantly, the full text was always available to all users, regardless of which learning style they selected. To keep experimental conditions controlled, the initial learning style choice could not be changed during the course of the study. Furthermore, participants using different learning styles were evenly distributed across all experimental conditions.

In order to investigate how malfunctions and explanations (or lack thereof) influence trust towards the system, we designed a within- and between-subjects study in which the system's accuracy and explainability were manipulated. The main study was preceded by a pre-test conducted in February 2021 and a pilot study. The pilot and the main study had the same number of texts and quizzes (i.e., seven).

3.1.2 Participants. The pilot study was conducted between April and July 2021. Over a period of seven weeks, each participants had a total of seven interaction sessions with PLANT, with one text and quiz per week. Participants were recruited from TU Wien. 75 participants took part in the pilot study, but only 13 participants completed it (2 female and 11 male). Their age ranged between 22 and 54 years old (M = 26, SD = 8,51). The highest educational degree completed by the participants was a general qualification for university entrance (46 %), bachelor's degree (46 %) and master's degree (1 %). The majority of participants were of Austrian nationality (85 %).

Participants for the main study were recruited through the online platform Probando [3] and were redirected to the PLANT website. Of the 205 participants who took part in the study, 171 completed it. Thus, the sample used for quantitative analysis consisted of 171 participants. In order to reduce the number of dropouts, in the main study, which was conducted between June and August of 2021, each participant had a total of seven interaction sessions with PLANT over a period of seven working days.

The majority of participants had Austrian nationality (72.5%) Participants' age ranged between 19 and 69 years old, with an average age of 29.3 years (M = 29; SD = 9.11). The majority of participants (71 %) identified as female The majority of participants had a general qualification for university entrance as their highest educational degree (51 %).

Before taking part in the study, participants were provided with information about the study and a consent form, which was approved by the Research Ethics Coordinator of TU Wien. After creating an account and logging in, participants were directed to the homepage of PLANT, where introductory information about PLANT, a timeline and assignments (quiz and questionnaire) were listed. Upon registration, they were asked to fill out a demographics questionnaire and learning style questionnaires. After submitting these questionnaires, PLANT suggested a learning style to each participant.

3.1.3 Experimental Conditions. Participants were then randomly assigned to one of the four experimental conditions. Hereby, we briefly describe each of them.

- **Correct with explanation (CwE)** PLANT provides correct recommendations throughout the entire study. From the beginning and throughout the study, the system allows participants to access a short explanatory description of how recommendations are generated.
- **Correct without explanation (CwoE)** PLANT provides correct recommendations throughout the entire study but does not offer any explanation concerning its inner workings.

[2] https://vark-learn.com/the-vark-questionnaire/, last accessed December 2021
[3] https://www.probando.io, last accessed September, 2021

- **Faulty with explanation (FwE)** PLANT initially provides three correct recommendations to let participants familiarize themselves with the system and to support trust formation. At the fourth interaction, the system provides a faulty recommendation (i.e., trust violation) and offers an explanation focused on the inaccuracy of one of the algorithms used by the system. The final three recommendations are again correct.
- **Faulty without explanation (FwoE)** PLANT initially provides three correct recommendations to let participants familiarize themselves with the system and to support trust formation. At the fourth interaction, the system provides a faulty recommendation (i.e., trust violation) and offers no explanation for the malfunction. The final three recommendations are again correct.

3.1.4 Procedure. Each participant was required to prepare for seven quizzes over the next seven working days. The texts that participants had to study were all related to emerging technologies and carefully selected by the team of researchers running the study. The order of the texts presented to participants across the four experimental treatments was fixed to mitigate the possibility of cheating (i.e., participants talking to each other about the previous quizzes). The order of the texts, respectively labeled from 'A' to 'G', plus 'X' and 'Y' (same text, but 'X' faulty, 'Y' correct) is reported in Table 1 below.

Table 1. Order of the texts in the different experimental treatments

Order	Participants
'A','B','C','X','D','E','F','G'	even userID
'A','D','E','X','F','B','C','G'	odd userID
'A','D','F','B','C','Y','E','G'	even userID
'A','B','D','F','E','Y','C','G'	odd userID

Each day over the following seven days, according to the provided timeline, a new text to study was made available on the homepage under the 'assignment' section. After studying the text, participants were asked to take a short quiz that consisted of five multiple-choice questions about the text. Participants could take the quiz whenever they wanted during that day. After clicking on the quiz, they only had five minutes to complete it. After each quiz, participants were asked to fill out a post-test questionnaire that contained questions about trust and satisfaction levels.

Upon completion of the seventh quiz, the study concluded with a final questionnaire that contained questions about participants' perceived trust in the system and perception of the system's usefulness. After finishing the study, they received an email informing them of the review and payment process and inviting them to participate in an online interview and focus group about their perception of the whole experience (participation in the interview and focus group were optional and unpaid).

As a token of gratitude for participants' time and support, all those who completed all the questionnaires (demographic & learning style, post-test, final questionnaire) received a fix payment of 35 Euro. Additionally, for each correct response to a quiz question, they received a bonus payment of 0.5 Euro (i.e. if a participant answered all five questions correctly in all seven quizzes, they received a bonus of 17.5 Euro, yielding a total compensation of 52.5 Euros.) However, participants in the pilot study received a participation certificate signed by the head of the research group. In addition, everyone who completed all questionnaires was entitled to participate in a lottery with ten prizes

worth 200 Euro each. The lottery drawing was a live online event conducted in July under the supervision of a member of the research ethics coordination team at TU Wien.

Eventually, participants were debriefed via email about the actual purpose of the study and the fact that the system was not actually autonomous and was operated by humans.

3.2 Measurements

Questionnaires. Trust perception was measured by means of an adapted version of the short, validated 'Trust Perception Scale-HRI', consisting of twelve items [54]. We used the short version as it is suitable for "trust measurement specific to measuring changes in trust over time, or during assessment with multiple trials" [54, p.214] and because it is specific to systems' functional capabilities. A sample item was "What % of the time did PLANT perform exactly as instructed". Three negatively worded items (Items 1.8, 1.10, 1.11) were reverse coded. Two items that directly and specifically referred to physically embodied robots were excluded from our questionnaire.

At the end of seven-day study, participants were asked to rate the trustworthiness of PLANT. Trustworthiness was measured by means of an adapted version of the 'Multi-Dimensional Measure of Trust' [40], which consists of 16 items divided into four groups (namely capable, reliable, ethical, sincere). Only the scale's wording was adapted to fit our specific use case. Participants were asked to report how closely they associated PLANT with each item on a five-point scale ranging from "strongly disagree" to "strongly agree". A sample item was "Predictable". Cronbach's alpha for the four sub-scales were: capable= 0.86, reliable= 0.78, ethical= 0.89, sincere= 0.88.

General risk propensity was measured with the 'General Risk Propensity Scale' (GRiPS), consisting of eight items [67]. A sample item was "Taking risks makes life more fun". Cronbach's alpha was 0.88. The participants indicated their level of agreement with statements on a five-point scale ranging from "strongly disagree" to "strongly agree".

As previously noted, participants were suggested a learning style based on their responses to the VARK Questionnaire Version 8.01 Finally, demographic information such as age, gender, highest educational degree, and country of citizenship was acquired.

Interviews and focus groups. We conducted 18 semi-structured interviews and a focus group discussion. The interviews focused on the following topics: the functionality and purpose of PLANT, the personalized learning styles, experiences concerning reliability, explanations and interpretability of PLANT. However, since the interviews were semi-structured, other topics emerged as well. The focus group concerned the same topic; however, there was a stronger emphasis on the explanations and malfunctions of PLANT, since the focus group allowed for fruitful discussions on these topics.

We conducted the interviews and focus groups online with the help of video conference software [4]. The data was collected in the form of audio recordings, which were subsequently transcribed using transcription software [5]. Both the audio recordings and transcripts were stored in a protected database at TU Wien. Only PLANT team members had access to this database. After the transcription of the interviews and focus group, we analysed the textual data with the help of the Atlas.ti [6] qualitative data analysis software. The analysis was conducted using a qualitative coding methodology in order to increase the validity of the findings and decrease bias. Furthermore, this analysis was conducted by two different people. Several meetings were organized in order to discuss the direction of the coding process. Furthermore, the functionalities of the qualitative data analysis software gave us a good overview of the major topics that played a role in the qualitative research.

[4]https://zoom.us, last accessed December, 2021
[5]https://www.otter.ai, last accessed December, 2021
[6]https://atlas.ti, last accessed December, 2021

4 RESULTS

4.1 Quantitative analysis

The software SPSS (Statistical Package for the Social Sciences) version 26 (IBM Corp, 2019) was used to conduct a series of one-way ANOVAs, t-tests, repeated measures ANOVAs and subsequent pairwise comparisons (Bonferroni corrections applied) to explore how trust develops over the course of repeated interactions and is restored after a violation.

According to Figure 5, the distribution of risk propensity was bi-modal, which implies two separate classes among the participants in the sample (M = 2.90, Md = 2.87, SD= 0.71).

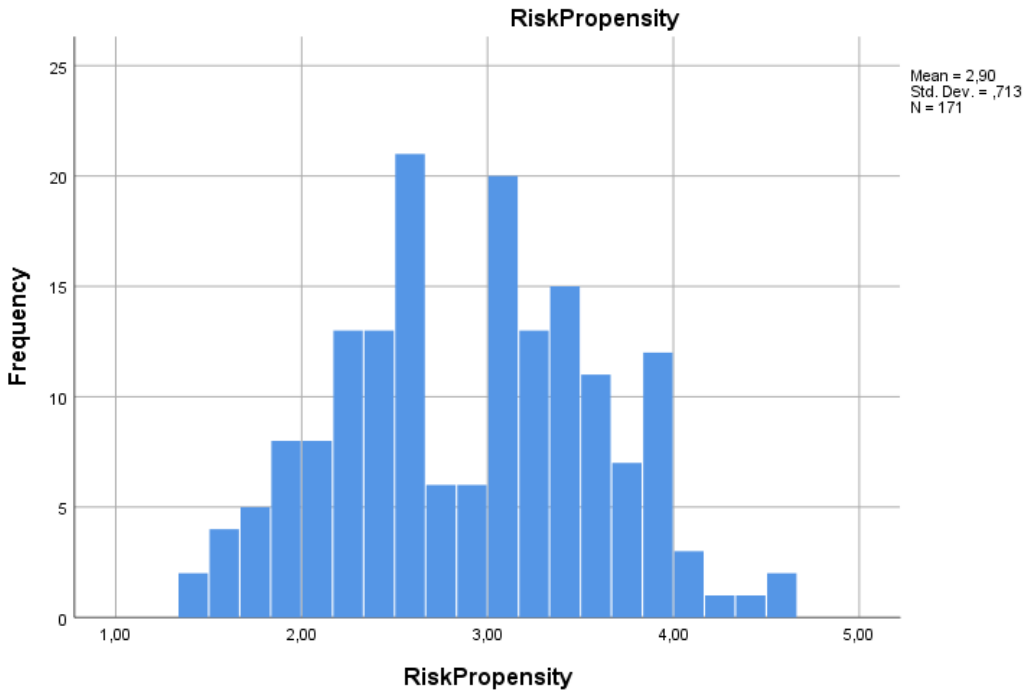


Fig. 5. Bimodal distribution of risk propensity

Table 2. Frequency distribution of treatment groups

Group	Frequency	Gender			Learning style			
		male	female	other	Kinesthetic	Visual	Auditory	Reading
CwE	37 (22%)	14	23	0	18	8	8	3
CwoE	52 (30%)	14	38	0	21	12	5	14
FwE	46 (27%)	8	37	1	21	8	5	12
FwoE	36 (21%)	12	24	0	18	10	6	2
Total	171 (100%)	48	122	1	78	38	24	31

After filling out the demographics and learning style questionnaires, participants were randomly assigned to one of the four groups. Table 2 shows the frequency distribution of the groups. As each group was nearly equal in size (52/36=1.44 < 1.5), the multivariate test results are fairly robust.

4.1.1 Initial trust perception. An independent samples t-test was conducted to compare the initial trust level (day 1) in groups with and without the explanation. There was no significant effect of the explanation on initial trust level, t(169) = -0.59, p = 0.56, even though both groups with the explanation (namely, correct, i.e., CwE and faulty, i.e., FwE) (M = 86.15, SD = 10.99) exhibited higher trust scores than the groups without the explanation (CwoE and FwoE) (M = 84.99, SD = 14.32). Thus, H1a is not supported. Moreover, no significant correlation was found between risk attitude and initial trust level (r= - 0.40, p = 0.60). Therefore, H1b is also not supported.

4.1.2 Trust development over time. To assess the effect of explanations on trust development over time, we looked at trust scores in both groups with no system malfunction (i.e. CwE and CwoE) (See Figure 6). A repeated measures ANOVA with a Greenhouse-Geisser correction indicated that mean trust level did not differ significantly during the 7 days within these two groups (CwE: F(3.82, 133.77) = 1.86, p = 0.12; CwoE: F(4.37, 222.94) = 2.03, p = 0.08). Thus, our results do not support H2.

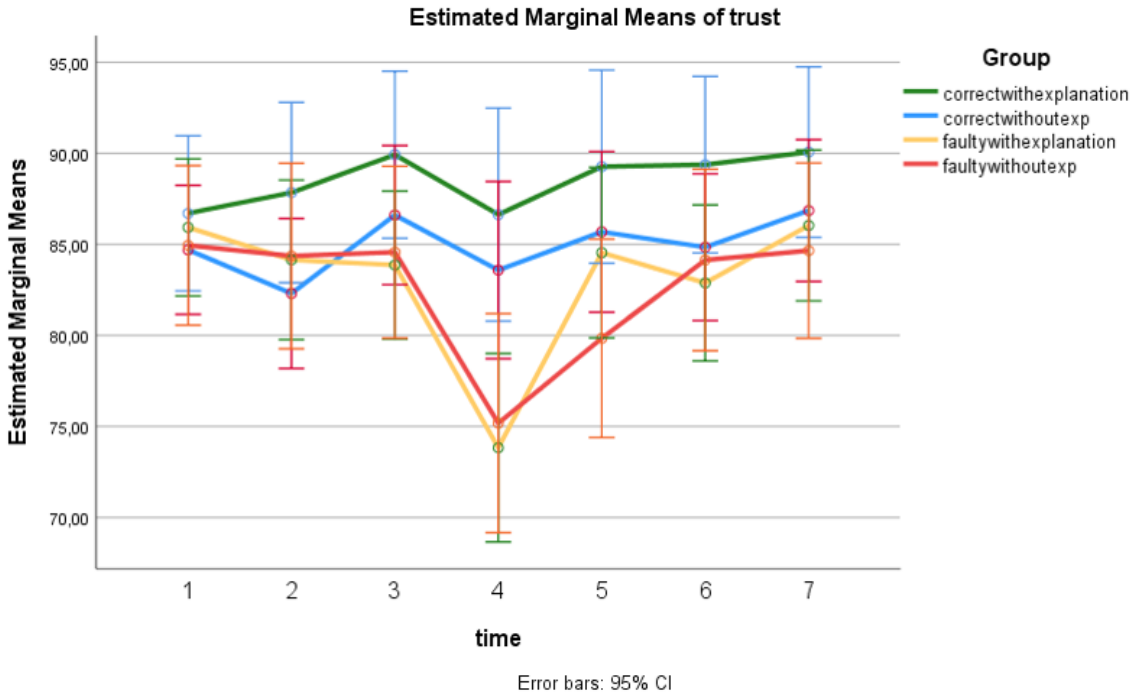


Fig. 6. Trust development in each group

4.1.3 Trust violation and restoration. Figure 6 shows how trust developed over time among all four groups. The trust level was lower on day 4 in groups with the malfunction (FwE: Mean = 73.84; FwoE: Mean = 75.19) compared to groups without the malfunction (CwE: Mean = 86.64; CwoE: Mean = 83.59). An independent t-test revealed that this difference among groups with and without the malfunction is significant (t(168) = 3.74, p < 0.01). Thus, H3a is supported.

On day 5, while descriptive statistics revealed that participants' trust was higher in the faulty with explanation group (FwE: Mean = 84.55) compared to the faulty without explanation group (FwoE: Mean = 79.85), an in-dependent t-test showed that this difference was not significant (t(79) = 1.02, p = 0.31) (see Table 3).

In the FwE group, a repeated measures ANOVA with a Greenhouse-Geisser correction determined a significant difference in trust level from day 4 to day 5 (F(1, 45) = 25,06, p < 0.001). Post hoc analysis with a Bonferroni adjustment revealed that trust level statistically significantly increased

Table 3. Mean scores for trust perception by groups (day 3-5)

Day	Group	Mean	SE	LB	UB
3	CwE	89.93	2.32	85.34	94.52
	CWoE	86.61	1.93	82.79	90,43
	FWE	83.87	2.06	79.81	87.93
	FWoE	84.57	2.39	79.85	89.30
4	CwE	86.64	2.96	80.79	92.49
	CwoE	83.59	2.46	78.72	88,46
	FwE	73.84	2.62	68.66	79.01
	FwoE	75.19	3.05	69.17	81.20
5	CwE	89.27	2.68	83.97	94.57
	CwoE	85.69	2.23	81.28	90.10
	FwE	84.55	2.37	79.86	89.23
	FwoE	79,85	2.76	74.40	85.30

(10.71 % CI, 6.40 to 15.02, $p < 0.001$). In contrast, we found no significant difference in trust level from day 4 to day 5 for the FwoE group ($F(1, 34) = 2{,}92$, $p = 0.10$). Thus, H3b is supported.

In light of the bi-modal distribution of risk propensity among participants, we classified the participants in two groups: i) risk averse with low risk propensity (mean < 2.90), and ii) risk tolerant with high risk propensity (mean ≥ 2.90). As Figures 7 and 8 show, the trust level in the faulty with explanation group improved significantly from day 4 to day 5 only among risk-averse individuals (14.07 % CI, 1.07 to 27.07, $p < 0.05$), not for the risk tolerant group (8.12 % CI, -0.09 to 16.34, $p = 0.06$). For the faulty without explanation group, the differences from day 4 to day 5 were non-significant for both risk-averse (5.29 % CI, -11.91 to 22.49, $p = 1.00$) and risk-tolerant individuals (3.95% CI, -5.04 to 12.95, $p = 1.00$). That implies that an explanation about a malfunction repaired trust to a greater extent among risk-averse compared to risk-tolerant participants. Taking all this together, H3c is supported, as risk propensity affected trust restoration.

Table 4. Mean scores for trust perception in the risk averse and risk tolerant groups (day 4 to day 5)

RP	Day	Group	Mean	SE	LB	UB
< 2.90 (Risk averse)	4	CwE	85.29	4.03	77.28	93.30
		CwoE	86.67	3.40	78.90	92,44
		FwE	74.46	4.03	66.45	82.47
		FwoE	76.03	4.24	67.59	84.47
	5	CwE	90.49	3.07	84.37	96.60
		CwoE	89.35	2.60	84.18	94.52
		FwE	88.54	3.07	82.42	94.65
		FwoE	81,32	3.24	74.87	87.77
≥ 2.90 (Risk tolerant)	4	CwE	88.33	4.46	79.45	97.21
		CwoE	81.16	3.64	73.91	88.41
		FwE	73.36	3.50	66.39	80.32
		FwoE	74.23	4.46	65.35	83.12
	5	CwE	87.75	4.52	78.76	96.74
		CwoE	81.42	3.69	74.08	88.76
		FwE	81.48	3.54	74.43	88.53
		FwoE	78.19	4.52	69.20	87.18

RP= Risk propensity

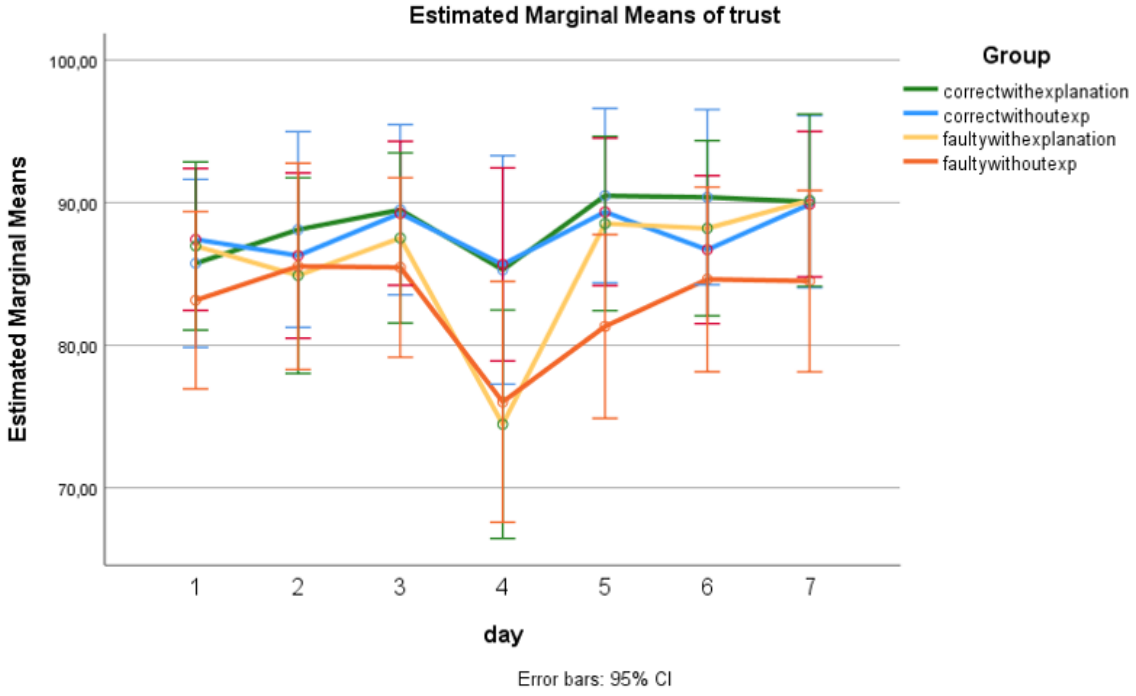


Fig. 7. Trust development in each group - risk averse

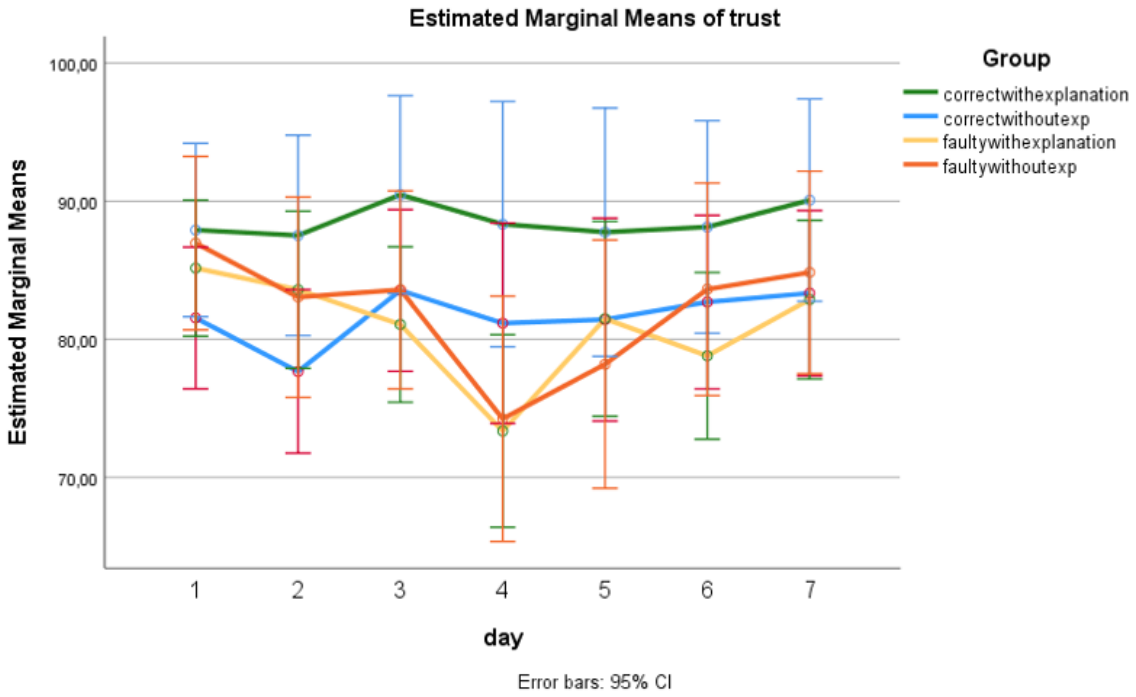


Fig. 8. Trust development in each group - risk tolerant

4.1.4 Trustworthiness and time spent studying. As described in the procedures section, participants were asked to rate the trustworthiness of PLANT at the end of study. Table 5 shows the mean, median and standard deviation of four different dimensions of trust: reliable, capable, ethical, and sincere. These four dimensions were organized into two broader trust concept:capacity trust (reliable, capable) and moral trust (ethical, sincere). A one-way ANOVA revealed that the ratings for capacity trust and moral trust did not statistically differ by groups (capacity trust: $F(3,167) = 1.59$, $p = 0.19$; moral trust: $F(3,164) = 0.65$, $p = 0.58$).

We also measured the time spent studying for each quiz. As shown in Figure 9, there was a downward trend in the average time spent studying with PLANT after day 5 in the FwoE group (day 4: M = 24.42; day 5: M = 24.28; day 6: M = 21.08; day 7: M = 19.64). However, in the FwE group, no increase or decrease was observed (day 4: M = 16.98; day 5: M = 15.93; day 6: M = 16.28; day 7: M = 16.26).

Table 5. Mean, median and standard deviation of multidimensional measure of trust by group

Trust dimension	Group	M	Median	SD
Capable	CwE	4.24	4.25	0.60
	CwoE	4.14	4.25	0.75
	FwE	4.13	4.00	0.62
	FwoE	3.96	4.00	0.80
Reliable	CwE	4.33	4.25	0.52
	CwoE	4.02	4.00	0.69
	FwE	4.04	4.00	0.70
	FwoE	4.04	4.00	0.58
Ethical	CwE	4.02	4.00	0.73
	CwoE	3.92	4.00	0.75
	FwE	3.99	4.00	0.75
	FwoE	3.85	3.75	0.69
Sincere	CwE	4.06	4.00	0.72
	CwoE	3.96	4.00	0.83
	FwE	3.94	3.75	0.67
	FwoE	3.78	3.87	0.68

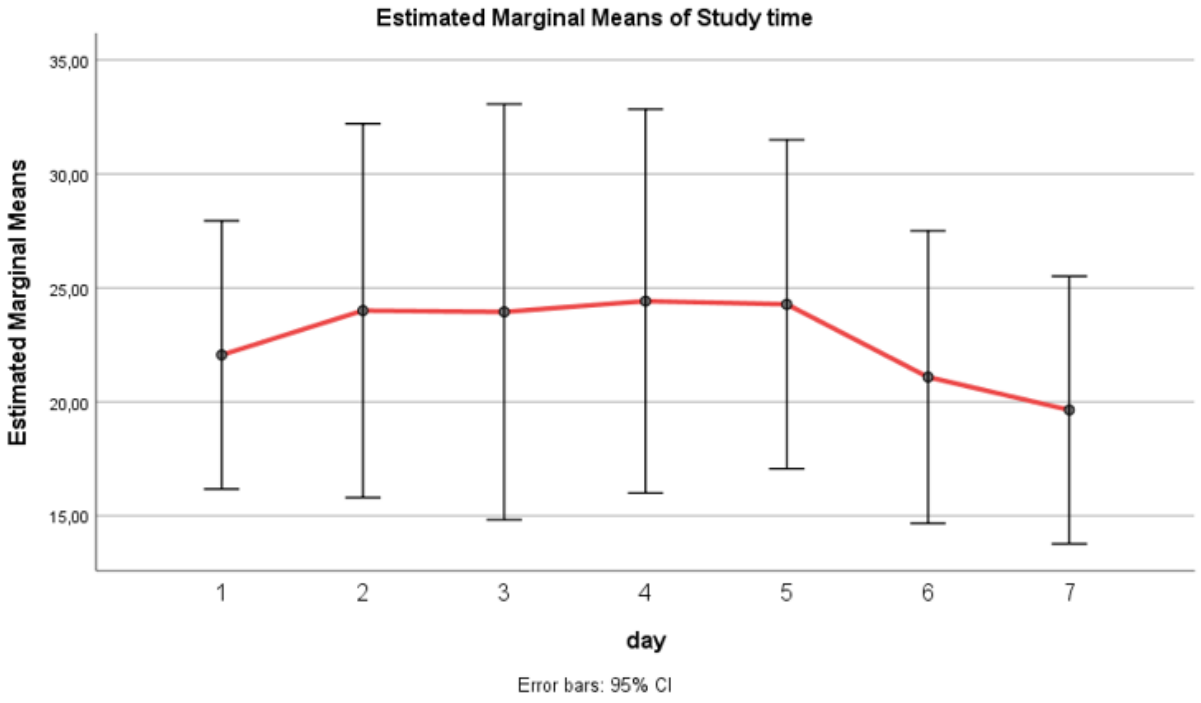


Fig. 9. Time spent studying in faulty without explanation group

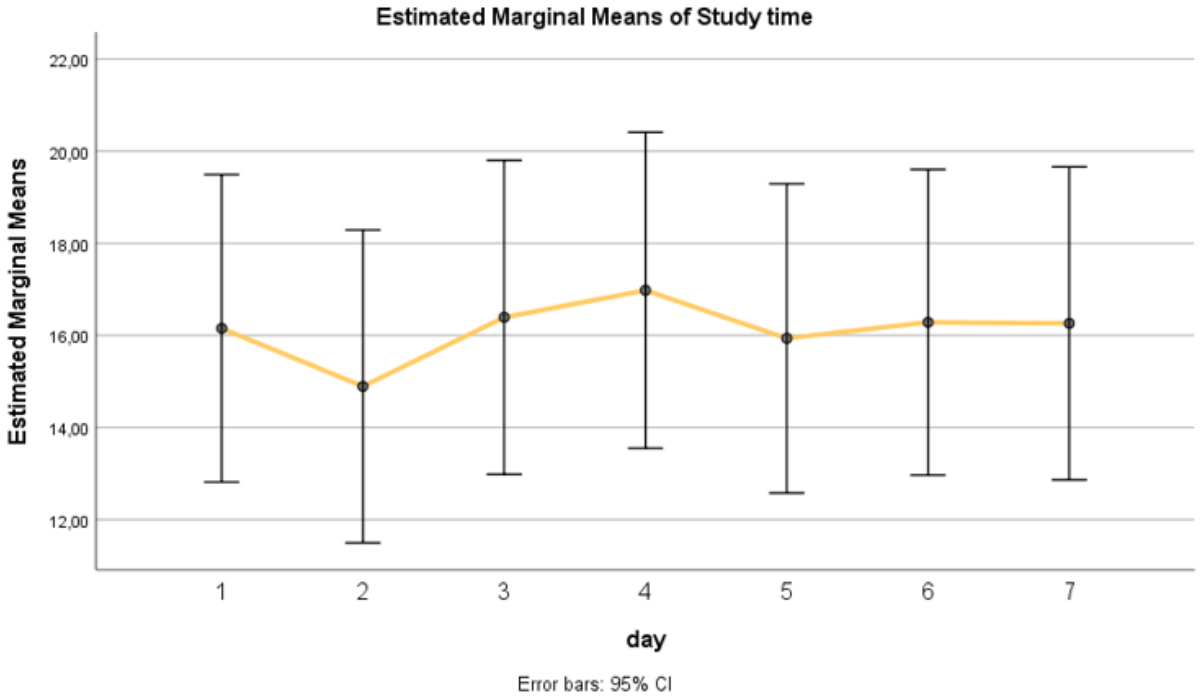


Fig. 10. Time spent studying in faulty with explanation group

4.2 Qualitative analysis

After finishing the study, participants received an email and were asked whether they wanted to provide us with further feedback and insights by participating in optional interviews and focus groups. As indicated above, we conducted 18 semi-structured interviews and one focus group discussion. The results from the qualitative research provided useful insights that supplemented the quantitative results. In what follows, we focus on three different major outcomes of the qualitative research, namely the perceived accuracy and reliability of PLANT, perceptions of the malfunctions of PLANT, and how participants experienced system transparency in terms of explanations.

4.2.1 PLANT's accuracy and reliability. First of all, as already indicated in the theoretical background section above, experiences regarding a system's accuracy and reliability are central components of the general process of building trust in a system. Participants generally provided constructive and positive feedback about their experiences with the assistance offered through PLANT´s recommendations. Their experience of PLANT's accuracy was a central element in that regard. Furthermore, perceived accuracy was also directly connected to their descriptions of PLANT's reliability. Naturally, both of these aspects are important for trust building and trust repair. Therefore, our qualitative research focuses less on proving the relation between experiences of accuracy and reliability and trust, but rather on gathering complementary insights and explanations about how this affected participants' trust in PLANT's assistive features.

A first topic that came to the fore concerning the experiences participants reported about the accuracy of PLANT was the participants' level of understanding about the output of PLANT's assistive features. That is to say, the more people compared the original text with the raw text input, the better their impression of PLANT's accuracy. Several participants explained that after reading the original text, they gained an increased understanding of PLANT's accuracy with regard to its assistance for the quizzes and thus also started to appreciate PLANT as a reliable assistant.

However, in some cases, this also led to a strong curiosity, which could in turn lead to certain suspicions about how this was done in an automated manner.

Furthermore, in this context, people used language expressing the importance of both accuracy and reliability as components of their individual aims in using PLANT. In that regard, the experience of accuracy and reliability was described as something relative to the user's needs. In our case, participants who were focused on receiving assistance for the test were mostly happy with the system's accuracy, even if they had read the full text beforehand. However, other participants were instead focused on getting a better understanding of the text content in general (and less focused on the test). Since texts can always be read in different ways, they would occasionally complain about the accuracy of the assistance.

Furthermore, something that came up in relation to the phenomenon of trust were participant's speculations about the role of PLANT's automated element. This is a recurring topic that will also be mentioned below. In terms of accuracy, an important outcome was that several participants expressed a desire to gain more insight into how PLANT could achieve this kind of accuracy. In other words, such people were looking for more explanations, regardless of whether the system malfunctioned or not. This provides an important lesson about the experience of trustworthiness in the sense that trust in the system is not just an outcome of the experience of accuracy itself, but is likely influenced by the larger context within which people interpret this accuracy. This could be particularly interesting in terms of achieving better explanations to improve the interpretability of assistive recommender systems, since curious users can be provided with clearer insights into how such systems function. To provide just a few examples, this can include information about the systems' developers, insights into the choices made during the development phase, and so on.

4.2.2 Perception of malfunctions. PLANT's malfunctions were a central element of the experiment and the subsequent quantitative analysis. Interestingly however, malfunctions were not always explicitly experienced as a prominent issue by the participants. In fact, even if PLANT provided a wrong recommendation and the participants were asked about this, they were not always bothered by the malfunction. Instead, some participants were rather forgiving, generally due to the fact that PLANT was framed as an automated system. The very notion that the system was fully automated served as an explanation for behavior perceived as a mistake from a human perspective. That is to say: the automated system did something that could be experienced as a malfunction by participants, but at the same time, some participants were implicitly aware that these kinds of malfunctions are normal in cases where humans and automated systems need to adjust to each other. This emphasis on two different types of intelligence, human versus artificial, can therefore be considered important for the way people interpret and judge such malfunctions.

4.2.3 Transparency through explanations. Finally, with regard to PLANT's explanations, several interesting insights emerged in the qualitative research. The focus group provided particularly interesting results here, where a discussion emerged regarding the general role of explanations in automated systems. First of all, the word "black box" played an important role in this context, as it was used to emphasize how it was still not clear to participants how exactly the mechanisms behind PLANT's assistance worked. Even though this was obviously related to the fact that this was a Wizard of Oz study, in hindsight, a good solution would have been to provide clear, additional insights into the way NLP systems develop text summaries.

Generalizing the statements above to assistive recommender systems, it is interesting to see that several participants exhibited strong curiosity regarding the explanations of how such systems work, as described above. When asking for more clarification about this curiosity in the focus group, a consensus emerged that different types of users should be provided with different kinds of explanations. That is, people agreed that users with different backgrounds are likely to be looking

for a divergent range of insights. For instance, the technical details behind the system might be interesting for a specific group of users, whereas others are likely to focus more strongly on the application's user-friendliness. It is therefore recommended that different explanations be able to be accessed through different channels. An example would be to implement explanations in the system itself, but also provide further explanations on the website about how such automated systems work, through social media channels or as part of personalized insights. Explainability in this sense can be seen as a term denoting a general tendency to provide explanations in many different ways.

Furthermore, a topic that came up in the qualitative research was the availability of explanations. Several participants argued that even if they would not access such explanations or insights into the data, they would prefer such explanations and insights to be available nevertheless. This is an important insight from the qualitative research, since it shows that even though explanations are not always accessed immediately (since people lack the time or the motivation to go through them), their very availability could help to create the impression of a system that is embedded in a larger framework of transparency.

Finally, in relation to such a larger framework, a topic that came up concerned the authority embedded in the explanation. Crucial here is to understand the way in which explanations are embedded in a larger array of expectations about a system's quality. Explanations can help to build and restore trust when they are seen as dependent on the perceived authority of the entity that provides the explanation. In other words, if the explanations are provided by people, institutions or companies that are already seen as reliable and transparent, participants reported that they would be much more likely to take the explanations for malfunctions or irregularities seriously.

5 DISCUSSION

Prior research has stressed the importance of longitudinal studies for understanding the development of trust over time [11, 22, 58]. To this end, our work extends previous work on trust dynamics in human interaction with AI-based systems. More specifically, our investigation sheds light on the combined effects of a recommender system's level of performance and explainability (or lack thereof) on people's trust over the course of repeated interactions.

Contrary to expectations and to part of the literature [4, 58], this study did not find a significant difference in terms of trust formation and continuous trust development between groups with and without explanations in the absence of any malfunctioning. Participants' propensity to take risks also did not significantly influence initial trust ratings. A possible explanation for these results regarding initial trust comes from the qualitative analysis. In fact, in the interviews and the focus group, the role of 'institutional cues' in the form of a 'concealed authority' behind the system's explanations emerged. It was not possible to determine with certainty how the researchers' authority influenced participants' perception of the system, particularly in terms of initial trust (e.g., in whether the system would behave benevolently). However, the fact that participants brought up the topic, specifically in relation to the reliability and transparency of such a 'concealed authority', corroborates the idea that environmental factors and external entities play a crucial role in determining people's initial perception of new technologies [4, 42].

Concerning continuous trust development, several studies (e.g. [9, 13]) suggest that explanations are unnecessary or even detrimental in some cases. For instance, [56] found that when a system's prediction are correct, providing more insights into the agent's decision-making process negatively affected trust. This perspective may offer a possible interpretation of our findings. When a system performs accurately over repeated interactions, people's expectation that the system will behave reliably consolidates [17, 45]. Accordingly, in such cases explanations become superfluous and do not necessarily increase people's trust in a system. Supporting this position, the qualitative analysis

demonstrated that some participants read the original text alongside PLANT's recommendations to check the system's accuracy, particularly during the first interactions. In other words, they did not immediately rely solely on the recommendations to prepare for the quizzes. Rather, they trusted them after checking them for themselves and after their accuracy was confirmed by the initial quizzes. On the other hand, the qualitative research also demonstrated that the availability of (different kinds of) explanations can be beneficial as a component of the overall contextual framework of a system's transparency and trustworthiness.

In this regard, we also found that not all participants who had the option accessed the explanations providing insights about how the system worked. While this may, in part, also explain why explanations did not yield higher trust in the initial phases, the qualitative analysis also indicated that even when participants did not access such explanations, their very presence added to the positive perception of the system. A possible interpretation for this is that explanations should not be forced upon users, particularly during the initial phases of an interaction and as long as a system performs accurately. Rather, they should be made available and it should be up to the users themselves to decide whether they need them or not [47]. Furthermore, it can be beneficial to provide a range of different kinds of (personalized) explanations.

With regards to trust violation and restoration, this study found that system malfunctions negatively influence trust ratings, as they are perceived as trust violations. These results corroborate findings from the literature about the negative effects of errors and malfunctions on trust [53, 62]. Interestingly, however, during the interviews and focus groups, some participants reported that they were not too negatively surprised by the system's faulty recommendation. Because they were aware of its non-human nature, they expressed forgiveness. The results from [12] suggest that mistakes and malfunctions that occur early on during an interaction affect trust more negatively than later ones. This may reconcile the apparent discrepancy in our findings, as the system provided a faulty recommendation only at the fourth interaction (i.e., after three correct ones). In turn, PLANT's initial accuracy may clarify why several participants were so tolerant towards the system's faulty recommendation, even though it still yielded significantly lower trust ratings.

Furthermore, we found that when the system made a faulty recommendation, transparency in the form of explanations generally led to significantly faster trust restoration. These results further support the notion of explanations as a trust restoration strategy [4, 16, 43, 51]. As previously discussed, if events such as mistakes and malfunctions are not dealt with, they are likely to undermine trust and negatively affect acceptance of technology [37]. In our study, this was reflected by the downward trend in the average time spent studying with PLANT in the faulty without explanation group after the system's faulty recommendation. The fact that participants did not stop using the system completely could be related to the reward scheme, as dropping out during the study would have led to no payment. Furthermore, findings from the qualitative analysis suggest that to maximize explanations' positive effect on trust, they should be tailored to the needs of specific groups of users. Specifically, comments from participants with different levels of familiarity and expertise with the technology suggest the need for different explanations that provide insights at different levels of complexity. This supports the results of studies showing how different types of explanations lead to different user reactions [63]. In addition, our qualitative research indicated that the feeling of familiarity with the system's features is likely to influence the perception of trustworthiness. These insights corroborate previous work suggesting that personalization of a system's features may have a positive impact on overall acceptance by increasing familiarity with the system [61].

Finally, another important finding was that the system's explanations about the malfunction were most effective in terms of restoring trust when participants were risk averse. This finding suggests that the success of trust repair through system transparency is tightly connected to a

person's risk attitudes and is consistent with [7], who found a negative effect of risk aversion on the relationship between perceived usefulness and trust.

6 CONCLUSIONS AND OUTLOOK

This paper analyzed how people's trust in a assistive recommender system evolves over time. Our main findings show that people perceive a system's malfunctioning, such as a faulty recommendation, as trust violation events, and it negatively affects trust ratings even after the system has proven reliable. To this extent, we observed a downward trend in time spent studying only in the faulty group without explanation. If, as Lomas et al. suggest [37], participants' use of the system can be considered an indicator of trust and technology acceptance, we can conclude that after a malfunction, users will start spending less time with the system if no explanation is provided.

Furthermore, while explanations did not yield greater trust at the beginning and throughout the interaction, they led to faster trust restoration after the mistake. This finding adds to the ongoing effort to understand the dynamics of people placing trust in AI-based systems as they relate to explainability. Another contribution of this study stems from its inclusion of risk aversion as a key user characteristic in relation to trust in the system. Determining subjects' risk attitude can help to optimally calibrate the level of transparency for human-AI interaction.

The results of the quantitative and qualitative investigations also suggested questions that are worth exploring in future work. For instance, reflecting the idea that 'institutional cues' may affect how people place trust in new technologies, an open question concerns how the researchers behind the study may have influenced participants' perception of the system (and its explanations) in terms of trustworthiness. Future work may wish to investigate in more detail the extent to which such external entities contribute to determining how people trust new AI-based systems, particularly in the initial phases of an interaction.

Another aspect that deserves more attention concerns explanations' personalization. Since AI-based systems are employed in a wide range of contexts, the needs of users will also likely vary noticeably in terms of depth, level of details and expertise. Accordingly, future work should address what AI-based system functionalities may offer such customization options, for instance in terms of explanations' 'interactivity' and 'multi-modality' [3]. Furthermore, future research would need to be conducted to understand the particularities of different types of transparency and explainability and how they affect trust in assistive systems.

The generalizability of these results is subject to certain limitations. For instance, this study did not find a difference in the ratings of capacity trust and moral trust between groups. This finding was unexpected, as it was speculated that groups who experienced a malfunction would rate PLANT lower than groups that did not experience a malfunction. This inconsistency may be due to the fact that the system's correct behavior after malfunction itself positively affected participants' trust that PLANT is capable of completing a task (i.e. capacity trust) and will not exploit the trustor's vulnerability (i.e. moral trust), as these constructs were assessed at the end of the study (after the seventh interaction) and not after the malfunction (on day 4). Perhaps a study design in which either these facets of trust are investigated immediately after the malfunction, or in which the system makes multiple mistakes throughout the interaction would yield higher consistency in trust ratings.

Finally, another possible limitation concerns participants' understanding of the causes of the faulty recommendation in relation to the quality of the explanations provided by the system. In fact, some participants noted that it was not clear how the system worked, even after the explanation. This perspective can be attributed in part to the fact that any explanation is only an imperfect approximation of an actual decision-making process [64], leaving room for misinterpretation. At the same time, it is important to acknowledge that an explanation which is not understood serves

little to no purpose in terms of proper trust calibration. Perhaps such a failure to understand the explanation of the system's inner workings was the reason why initial trust was marginally higher in the groups with the explanation. A study designed to explicitly assess how people understand different types of explanations (e.g., at different levels of depth and complexity) would help to shed light on the dynamics of understanding explanations.

ACKNOWLEDGMENTS

The authors would like to thank Georg Bixa and Reinhard Grabler for their support in developing PLANT. Funding for this study was contributed by Mercedes-Benz AG.

REFERENCES

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

[2] Ighoyota Ben Ajenaghughrure, Sonia Claudia da Costa Sousa, and David Lamas. 2020. Risk and Trust in artificial intelligence technologies: A case study of Autonomous Vehicles. In *Proceedings of the 13th International Conference on Human System Interaction (HSI)*. IEEE, 118–123.

[3] Kamran Alipour, Jurgen P Schulze, Yi Yao, Avi Ziskind, and Giedrius Burachas. 2020. A Study on Multimodal and Interactive Explanations for Visual Question Answering. *arXiv preprint arXiv:2003.00431* (2020).

[4] Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T Powers, et al. 2018. Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine* 37, 4 (2018), 76–83.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[6] Anthony L Baker, Elizabeth K Phillips, Daniel Ullman, and Joseph R Keebler. 2018. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 4 (2018), 1–30.

[7] Nesrine Ben Amor and Imène Ben Yahia. 2021. Investigating Blockchain Technology Effects on Online Platforms Transactions: Do Risk Aversion and Technophilia Matter? *Journal of Internet Commerce* (2021), 1–26.

[8] 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 (HFES Annual)*, Vol. 60. SAGE Publications Sage CA: Los Angeles, CA, 841–845.

[9] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction* 18, 5 (2008), 455–496.

[10] 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*.

[11] 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.

[12] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 251–258.

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

[14] Na Du, Jacob Haspiel, Qiaoning Zhang, Dawn Tilbury, Anuj K Pradhan, X Jessie Yang, and Lionel P Robert Jr. 2019. Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation research part C: emerging technologies* 104 (2019), 428–442.

[15] 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.

[16] AR 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* 22, 1 (2007), 4–24.

[17] Fabio Fossa. 2019. "I don't trust you, you faker!" On Trust, Reliance, and Artificial Agency. *Teoria* 39, 1 (2019), 63–80.

[18] Christopher P Furner, John R Drake, Robert Zinko, and Eric Kisling. 2022. Online review antecedents of trust, purchase, and recommendation intention: A simulation-based experiment for hotels and AirBnBs. *Journal of Internet Commerce* 21, 1 (2022), 79–103.

[19] Diego Gambetta. 2000. Can we trust trust? In *Trust: Making and Breaking Cooperative Relations*. Department of Sociology, University of Oxford, Chapter 13, 213–237.

[20] Yingqiang Ge, Shuchang Liu, Zuohui Fu, Juntao Tan, Zelong Li, Shuyuan Xu, Yunqi Li, Yikun Xian, and Yongfeng Zhang. 2022. A survey on trustworthy recommender systems. *arXiv preprint arXiv:2207.12515* (2022).

[21] Harjinder Gill, Kathleen Boies, Joan E Finegan, and Jeffrey McNally. 2005. Antecedents of trust: Establishing a boundary condition for the relation between propensity to trust and intention to trust. *Journal of business and psychology* 19, 3 (2005), 287–302.

[22] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[23] Hani Hagras. 2018. Toward human-understandable, explainable AI. *Computer* 51, 9 (2018), 28–36.

[24] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[25] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work (CSCW)*. 241–250.

[26] Kevin A Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[27] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st international conference on intelligent user interfaces (IUI)*. ACM, 164–168.

[28] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 624–635.

[29] Alexandra D Kaplan, Theresa T Kessler, J Christopher Brill, and PA Hancock. 2021. Trust in artificial intelligence: Meta-analytic findings. *Human Factors* (2021).

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

[31] Sherrie YX Komiak and Izak Benbasat. 2006. The effects of personalizaion and familiarity on trust and adoption of recommendation agents. *MIS Quarterly* 30, 4 (2006), 941–960.

[32] 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.

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

[34] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2022. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139 (2022), 1–18.

[35] Xin Li, Traci J Hess, and Joseph S Valacich. 2008. Why do we trust new technology? A study of initial trust formation with organizational information systems. *The Journal of Strategic Information Systems* 17, 1 (2008), 39–71.

[36] Steven Lockey, Nicole Gillespie, Daniel Holm, and Ida Asadi Someh. 2021. A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. In *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*. Hawaii International Conference on System Sciences, 5463–5472.

[37] Meghann Lomas, Robert Chevalier, Ernest Vincent Cross, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 187–188.

[38] Niklas Luhmann. 2000. Familiarity, confidence, trust: Problems and alternatives. *Trust: Making and breaking cooperative relations* 6, 1 (2000), 94–107.

[39] Fergus Lyon, Guido Möllering, and Mark NK Saunders. 2015. Introduction. Researching trust: The ongoing challenge of matching objectives and methods. In *Handbook of Research Methods on Trust: Second Edition*. Edward Elgar Publishing Ltd., 1–22.

[40] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*. Elsevier, 3–25.

[41] Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal* 38, 1 (1995), 24–59.

[42] D Harrison McKnight, Larry L Cummings, and Norman L Chervany. 1998. Initial trust formation in new organizational relationships. *Academy of Management review* 23, 3 (1998), 473–490.

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

[44] Hugo Neri and Fabio Cozman. 2020. The role of experts in the public perception of risk of artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 663–673.

[45] Onora O'neill. 2002. *Autonomy and trust in bioethics*. Cambridge University Press.

[46] Guglielmo Papagni, Jesse de Pagter, Setareh Zafari, Michael Filzmoser, and Sabine T Koeszegi. 2022. Artificial agents' explainability to support trust: considerations on timing and context. *AI & SOCIETY* (2022), 1–14.

[47] Guglielmo Papagni and Sabine Koeszegi. 2020. Understandable and trustworthy explainable robots: A sensemaking perspective. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2020), 13–30.

[48] LeeAnn Perkins, Janet E Miller, Ali Hashemi, and Gary Burns. 2010. Designing for human-centered systems: Situational risk as a factor of trust in automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting (HFES Annual)*, Vol. 54. SAGE Publications Sage CA: Los Angeles, CA, 2130–2134.

[49] Pearl Pu and Li Chen. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20, 6 (2007), 542–556.

[50] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM international conference on knowledge discovery and data mining (SIGKDD)*. 1135–1144.

[51] 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.

[52] Julian B Rotter. 1971. Generalized expectancies for interpersonal trust. *American psychologist* 26, 5 (1971), 443.

[53] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1–8.

[54] Kristin E Schaefer. 2016. Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In *Robust Intelligence and Trust in Autonomous Systems*. Springer, 191–218.

[55] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58, 3 (2016), 377–400.

[56] Philipp Schmidt, Felix Biessmann, and Timm Teubner. 2020. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 29, 4 (2020), 260–278.

[57] F David Schoorman, Roger C Mayer, and James H Davis. 2007. An integrative model of organizational trust: Past, present, and future. *Academy of Management review* 32, 2 (2007), 344–354.

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

[59] Jeffry A Simpson. 2007. Foundations of interpersonal trust. *Social psychology: Handbook of basic principles* 2 (2007), 587–607.

[60] 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.

[61] JaYoung Sung, Rebecca E Grinter, and Henrik I Christensen. 2009. "Pimp My Roomba" designing for personalization. In *Proceedings of the 27th Conference on Human Factors in Computing Systems (SIGCHI)*. 193–196.

[62] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M Powers, Clare Dixon, and Myrthe L Tielman. 2020. Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 3–12.

[63] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 109–116.

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

[65] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces (IUI*. 318–328.

[66] Jessie X Yang, Vaibhav V Unhelkar, Kevin Li, and Julie A Shah. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 408–416.

[67] Don C Zhang, Scott Highhouse, and Christopher D Nye. 2019. Development and validation of the general risk propensity scale (GRiPS). *Journal of Behavioral Decision Making* 32, 2 (2019), 152–167.

[68] Lynne G Zucker. 1987. Institutional theories of organization. *Annual review of sociology* 13, 1 (1987), 443–464.

Received 07 December 2022; revised 20xx; accepted 20xx

CHAPTER 5

Paper4

Minds and Machines (2021) 31:505–534
https://doi.org/10.1007/s11023-021-09567-6

ORIGINAL ARTICLE

A Pragmatic Approach to the Intentional Stance Semantic, Empirical and Ethical Considerations for the Design of Artificial Agents

Guglielmo Papagni[1] · **Sabine Koeszegi**[1]

Received: 31 December 2020 / Accepted: 5 July 2021 / Published online: 26 July 2021

Abstract

Artificial agents are progressively becoming more present in everyday-life situations and more sophisticated in their interaction affordances. In some specific cases, like Google Duplex, GPT-3 bots or Deep Mind's AlphaGo Zero, their capabilities reach or exceed human levels. The use contexts of everyday life necessitate making such agents understandable by laypeople. At the same time, displaying human levels of social behavior has kindled the debate over the adoption of Dennett's 'intentional stance'. By means of a comparative analysis of the literature on robots and virtual agents, we defend the thesis that approaching these artificial agents 'as if' they had intentions and forms of social, goal-oriented rationality is the only way to deal with their complexity on a daily base. Specifically, we claim that this is the only viable strategy for non-expert users to understand, predict and perhaps learn from artificial agents' behavior in everyday social contexts. Furthermore, we argue that as long as agents are transparent about their design principles and functionality, attributing intentions to their actions is not only essential, but also ethical. Additionally, we propose design guidelines inspired by the debate over the adoption of the intentional stance.

Keywords Intentional stance · Robots · Virtual agents · Ethics

1 Introduction

Artificial agents (i.e., physical robots, virtual agents and embodied virtual agents) capable of taking decisions autonomously are being employed in a growing number of fields. Several applications of such agents directly influence everyday life for a growing number of people, most of whom can be considered non-expert end-users.

✉ Guglielmo Papagni
guglielmo.papagni@tuwien.ac.at

[1] Institut für Managementwissenschaften, TU Wien, Theresianumgasse 27, Vienna 1040, Austria

To properly address the challenges posed by the integration and acceptance of artificial agents within society, coordinating the contributions of various disciplinary fields is necessary and a challenge within the challenge.

Being able to correctly understand and predict the behavior of artificial agents is necessary for the development of trustworthy relationships (De Graaf & Malle, 2017; Miller, 2019). In this regard, one fundamental feature addressed by the ongoing interdisciplinary efforts concerns whether people consider artificial agents' actions and decisions as intentional. Contributions on this topic are embedded into the broader framework of discussion over the attribution of anthropomorphic and social traits to artificial agents (Nass et al., 1994; Reeves & Nass, 1996; Dreyfus et al., 2000; Nass & Moon, 2000). However, in recent times, the idea that artificial agents' actions might be interpreted as intentional is emerging as a structured and semi-autonomous debate. A significant share of research on this issue is informed by Daniel Dennett's concepts of 'intentional systems' and 'intentional stance' (Dennett, 1971, 1981, 1989). The latter represents a strategy that people can adopt to make sense of and predict the behavior of complex rational (or intentional) systems, being them human agents or machines (Dennett, 1988). As the inner complexity of artificial agents grows, researchers in the fields of human–robot and human–computer interaction (HRI and HCI respectively) are investigating from multiple perspectives how people ascribe intentions to artificial agents.

Objections to the concept of intentional stance have been raised over the years. We recognize part of this criticism, specifically that which targets Dennett's commitment to a behaviorist perspective as reasonable. However, some of these objections tend to conflate biological definitions of intentionality with the attribution of intentions to artificial entities (Thellman et al., 2017), hence overshadowing what we consider Dennett's most relevant contribution. Our aim is to build upon the originality of his intuition without committing to Dennett's behaviorism. First, this paper provides a critical analysis of Dennett's concept, particularly in light of definitions of intentionality and some of the main objections. Moving beyond Dennett's position, we emphasize that the ascription of intentionality does not necessarily imply artificial agents to have genuine mental states in the human sense. Rather, we claim that the main quality of this strategy is to help people to manage social interactions with artificial agents, and that research should focus on how to maximize this positive aspect.

Furthermore, some scholars note a pressing lack of comparative analysis (Thellman et al., 2017). We argue that only through such a systematic approach it is possible to trace the point of origin of the attribution of intentions to artificial agents. Therefore, the paper analyzes relevant cases from the HRI and HCI experimental literature. From this comparative perspective, we discuss and refine Dennett's idea that complex rational behavior is the spark that ignites the process of intention ascription.

The idea of deceptive anthropomorphism offers another criticism of features that trigger the attribution of intention. The last section investigates ethical implications of the concept, particularly in light of recent calls for machine transparency and the risks of 'deceptive design'. Our claim is that the implementation of features that make artificial agents' behavior 'seemingly intentional' is not only ethical, but can

be desirable, as it can positively contribute to the overall quality of social interactions. However, we also identify a key condition for this assumption to hold. Users must be made aware of the nature of the agent they are interacting with before the interaction unfolds. Failing to fulfil this condition, as our examples show, can have negative consequences that could jeopardize the successful societal integration of artificial agents.

2 Critical Approach to the Intentional Stance

Each of the stances discussed by Dennett represents a strategy for understanding and predicting how certain entities work (Dennett, 1981, 1988, 1989). People adopt the physical stance to make sense of and predict the future behavior of certain systems, via their knowledge of physical laws. The design stance allows predictions to be made based on the assumption that systems work as they are meant to by design. In certain cases, however, these two strategies do not suffice. In particular, it may not be possible, let alone practical, to predict how rational systems (or agents) will behave based on the two previous strategies. To address this limitation, Dennett introduces the intentional stance. This is a predictive tool that relies on the assumption that rational systems will behave in accordance with their intentions, beliefs and desires in order to achieve specific goals (Dennett, 1981, 1988, 1989).

Dennett is neither the only one, nor the first to refer to artificial agents in "anthropomorphic" or social terms. One of the most relevant contributions in this direction is represented by the pioneering study from Heider and Simmel on the attribution of social meanings to the motion of geometric shapes (Heider & Simmel, 1944). Ever since the initial establishment of computing-related disciplines, researchers have spent significant efforts in trying either to make artificial agents appear and behave more like humans, or to explain why people tend to adopt social interpretative frameworks to understand and predict artificial agents' actions (Caporael, 1986; Nass et al., 1994; Breazeal & Scassellati, 1999; Breazeal, 2002).

2.1 Complexity of Intentional Systems

After this initial contextualization of his work, it is important to note that the reason why Dennett includes (certain types of) artificial agents as targets of the intentional stance lies in the complexity of these systems. In fact, referring to a chess-playing computer, Dennett says that such systems are "practically inaccessible to prediction from either the design stance or the physical stance; they have become too complex for even their own designers to view from the design stance." Therefore, one assumes "that the computer will 'choose' the most rational move" (Dennett, 1981, p. 5). In other words, Dennett emphasizes the idea that treating certain types of agents 'as if' they had intentions might in some cases be the only fruitful strategy to understand and predict their behavior. In fact, he continues, "when one can no longer hope to beat the machine by utilizing one's knowledge of physics or programming to anticipate its responses, one might still be able to avoid defeat by

treating the machine rather like an intelligent human opponent" (Dennett, 1981, p. 5). Again, elsewhere he clarifies that we adopt the intentional stance because "it gives us predictive power we can get by no other method" (Dennett, 1997, p. 66). At the same time, he also points out the difference between "those intentional systems that really have beliefs and desires from those we may find it handy to treat as if they had beliefs and desires" (Dennett, 1997, p. 66).

However, other positions expressed by Dennett on the topic contribute to further articulating the debate and give rise to some of the main critiques. In fact, in several occasions Dennett remarks the fact that if one were to infer and attribute any mental states to another agent, doing so through the analysis of the observable behavior of the agent would be the only way to go (Dennett, 1991, 1993). Dennett claims that there is no ineffable quality of the mind and that mental states can be discerned through the recognition and analysis of behavioral patterns. Furthermore, he argues that not only robots and artificial agents can be, in principle, referred to as "philosophical zombies", but that for the concerns of consciousness and mental states, everyone is such a zombie. This fictional entity is something that is functionally, i.e., behaviorally identical to a human being, but which lacks any form of actual consciousness (Dennett, 1993, 1995).

For the purposes of this paper, such considerations represent the problematic node in Dennett's framework. It is problematic because it implies the commitment to forms of "behavioral realism". Hence, from Dennett's perspective, either both humans and artificial agents can be considered conscious, at least as long as they behave in a qualitatively comparable way, or neither of them should. Not only that, the fact that Dennett subscribes to such positions triggers several critiques, so that the debate takes, at least partially, the direction of a diatribe on the overlapping of genuine intentionality (and mental states more in general) and the attribution of intentions to artificial agents (Thellman et al., 2017). The reason for this lies in the fact that, as Dennett poses that perceiving patterns of intentionality in the behavior of an agent corresponds to saying that those patterns are the one and only real thing, the only possible consequence is to match humans and artificial agents under the sign of the intentional stance. To clarify this fundamental point, the next paragraphs briefly discuss what in the literature is referred to as genuine intentionality, to then focus on some of the main critiques proposed against Dennett's arguments.

2.2 Biological Intentionality and Objections to the Intentional Stance

John Searle, for instance, refers to intentionality as a feature of an evolution-based mind that allows people to relate in the first place to each other, but also to the environment. "My subjective states relate me to the rest of the world, and the general name of that relationship is 'intentionality.' These subjective states include beliefs and desires, intentions and perceptions [...] 'Intentionality,' to repeat, is the general term for all the various forms by which the mind can be directed at, or be about, or of, objects and states of affairs in the world." (Searle, 1980, p. 85) Similarly, other philosophical definitions emphasize aspects of mind's relatedness to the world

(Jacob, 2019) and one's mental states as a function of their goals and aims (Miller, 2019).

Consequently, it might sound reasonable that in order to interpret and predict the behavior of even sophisticated devices such as robots, or conversational agents, it would be enough to know the purpose behind their design. As artificial entities, they are not endowed with the evolution-based features that contribute to the emergence of biological intentionality. Hence, from this perspective, it would not make sense to adopt the intentional stance when interacting with such machines. However, as it was previously noted, Dennett explicitly refers also to artificial agents in his formulations of the intentional stance and of intentional systems.

One of the shortcomings of Dennett's theory is addressed as the "ideal rationality of intentional systems". To this extent, it is argued that intentional biological agents do not always behave in full accordance with the ideal rationality implied by Stich (1985). Indeed, notes Stich, irrationality is a cornerstone of human behavior (Stich, 1985). If an intentional agent in Dennett's terms is one that always acts rationally, then intentionality and rationality necessarily go together, and if one acts irrationally, one cannot be an intentional agent, argues Stich. His main point is that this is not a valid argument to say that if one's behavior is not fully rational, then no intentions, beliefs etc. can be attributed.

Another line of argument directly targets Dennett's 'behaviorist' positions concerning intentionality and intelligence. What is criticized is the idea that if something behaves 'as if' it were intelligent (e.g., having intentions), then it should be considered as such, precisely because intelligence can be identified only through manifest behavior (Dennett, 1995; Danaher, 2020). Block makes the point that human-like behavior is not sufficient to characterize an agent as having human-like intelligence because, as a matter of fact, such behavior does not mirror actual mental states or intelligence. Rather, it is merely the manifestation of its programmers' intelligence (Block, 1981). Such machines, continues Block, lack "the kind of 'richness' of information processing requisite for intelligence" (Block, 1981, p. 28). In a similar fashion, Slors notes that Dennett never clarifies what the adoption of the intentional stance means without referring to intentions and hence winding up in a kind of circular argument (Slors, 1996).

At this point, one might note how the debate tends to be oriented towards a resolution of the contrast between genuine conceptions of intentionality (and intelligence), and the adoption of the intentional stance. In other words, it tends to relinquish the pragmatic usefulness of attributing intentions to artificial agents, as they can only display something that looks like intentionality, but lack the very substratum, processes and semantic richness that make up genuine intentionality. Whereas this might partially be Dennett's own fault in light of his behaviorism, we argue that this conflating attitude is itself part of the problem (Thellman et al., 2017).

If one aims to move beyond Dennett's behaviorism, why should a property unique to the mind be attributed to artefacts without minds? Furthermore, if certain types of machines, i.e., artificial agents, can be treated as if they were intentional, where should the line be drawn between them and devices whose behavior can be predicted only as a function of their design? And how is this boundary (if one exists) defined? These are the main questions addressed in this paper. Arguably, their relevance is

not merely philosophical. Since it is believed that artificial agents will play a central role in our society, it is fundamental to figure out appropriate design strategies to improve interactions with them and avoid ethically dangerous trends.

Based on the previous considerations, we argue that, for the purpose of social interactions with artificial agents, the usefulness of Dennett's proposal should not concern whether a machine could have genuine intentions. As the above mentioned critiques have shown, supporting the intrinsic behaviorism of the intentional stance might create more problems than it solves.

2.3 Alternative Semantics, Alternative Approach

In line with the position expressed by Thellman and Ziemke, we claim that rather than focusing on and committing to hard-to-prove ontological statements about the nature of mental states, the attention should be shifted towards the idea that, from a user perspective, treating a sophisticated agent 'as if' it were intentional might be the most appropriate strategy (if not the only one available). Attributing intentions to machines should be more about the mental states of the one doing the ascribing rather than the mental states (or lack thereof) of the machine itself (Thellman & Ziemke, 2019). To this extent, recalling Searle's definition of intentionality, Thellman and colleagues note that, in addition to reading intentionality as a function of relatedness (of subjective states to the world), Searle also refers to recognizing others as intentional agents as fundamental to predicting how they will behave (Thellman et al., 2017). Such a complementary (in Searle's terms) perspective reflects what we claim to be the most fruitful aspect of Dennett's formulation. People attribute intentions not necessarily or not exclusively to recognize conspecifics, but rather to understand and predict how people (and agents) behave (Dennett, 1988). This seems to reflect "folk-psychological" definitions which consider intentionality not only objectively (i.e., biologically) but also as a social construct that functions as a 'tool' to ease social interactions and thus also to make sense of or predict the behavior of sophisticated artificial agents (Dennett, 1988; Malle & Knobe, 1997).

In support of this idea, studies suggest that the general attitude to anthropomorphize artificial agents might be the default approach that people adopt, as a socio-cognitive construct, when they recognize human-like patterns in other agents' (human or non-human) behavior (Caporael, 1986; Nass et al., 1994; Caporael & Heyes, 1997; Breazeal & Scassellati, 1999; Nass & Moon, 2000). Within this perspective, the attribution of mental states to other agents emerges as an automatic, bottom-up process caused by the activation of brain areas responsible for social cognition as a response to the perception of human-like patterns and traits (Buckner et al., 2008; Looser & Wheatley, 2010; Spunt et al., 2015). Therefore, while it is true that such mechanisms are rooted in social cognition processes that humans developed in order to interact with each other, they are available also when interacting with artificial agents. At the same time, recognizing patterns does not necessarily imply believing that they reflect the same mental states, or any mental state at all.

Hence, we claim that for the consideration of the intentional stance to be fruitful, it should be interpreted as a strategy that people adopt, consciously or not, to

navigate the world of social interactions with other rational agents, human or artificial. To avoid the risk of conflation of the intentional stance with biological definitions of intentionality, ontological statements and commitment about artificial agents' hard-to-prove mental states should be left aside accordingly. As Breazeal states, referring to the Kismet robot, people treat it "as if it were a socially aware creature with thoughts, intents, desires, and feelings. Believability is the goal. Realism is not necessary" (Breazeal, 2002, p. 52).

Therefore, we suggest that perhaps an alternative, machine-specific semantic approach is more appropriate. One might argue that people attribute or infer intentions to other humans as well, so that attributing intentions to machines does not involve any different process. In a way, such overlap cannot be avoided, as the mental processes involved in the persons doing the ascribing are qualitatively the same, with the only difference being one of quantity. However, in light of the previous discussion about distancing from Dennett's behaviorism, what does differ qualitatively is how biological intentions and machine seemingly intentional behavior are generated. Whether or not users attribute intentions to such agents is supported and, to a certain extent, made possible by certain design strategies (as suggested in Wiese et al. (2017)).

Therefore, we claim that the emphasis should be put on the artificial and implemented nature of the features that trigger the ascription of intention in order to significantly mark the difference to biological instances. For instance, one could adopt formulations like artificial agents' 'seemingly intentional behavior'. This alternative approach should be accordingly interpreted as a pragmatic measure for researchers to work with, rather than as a specific ontological declaration. Following this premise, this strategy aims to describe the visible result, in terms of resembling intentional behavior, of specific implemented features. Since the ultimate goal of this discussion is to improve the quality of the interactions users undertake with artificial agents, if an agent's behavior appears to be intentional, we sustain that it should be possible to describe it in such terms without risking to bring out the implications of the previously discussed trend of 'conflating notions'.

On the side of users' experience, we deem the use of terms such as 'attributed' or 'ascribed' intentionality more appropriate than others like 'perceived'. 'Perceiving' intentions recalls the idea of the perceptual apparatus that people are endowed with, which collects, processes and reconstructs data from the surroundings (Malle, 2011). For instance, when an event occurs, the observable changes in the environment are perceived by our senses, recorded and processed. When it comes to social perception, perceptual information is combined so that people form impressions of each other and base their mutual judgement. Depending on the type of event or action that is perceived and processed, different types of causes (or reasons) are attributed as the result of a deliberative process (Parkinson, 2012). While the idea is in principle the same for both "object perception" and "person perception" (in which case intentions might be involved), the latter situation poses a more complex challenge, as the data and possible causes to be analyzed are more nuanced (Malle, 2011). It must be noted that such processes might not always occur on a fully conscious level. Rather, we might expect that the more people engage in relationships with artificial agents, the more specific mental models will be activated subconsciously, as

such interactions start belonging to implicit social cognitive processes (Evans & Stanovich, 2013). Such considerations aim to highlight the mental states of the people interacting with the agent. They emphasize that people resort, consciously or not, to the specific strategy of attributing intentionality in order to understand and predict the behavior of complex social machines, when other alternatives fail or cannot apply. As such, one can refer to the ascription of intentions without mentioning beliefs, desires and intentions themselves, as Slors argues (Slors, 1996).

Furthermore, we note how formulations such as 'simulated intentionality' might be proposed, in line with the idea of 'simulated social interactions' as opposed to 'fictional interactions' formulated in Seibt (2017). While it might be true that a given implemented feature aims to simulate intentional human behavior, labeling an interaction or implemented characteristics as 'simulated' suggests a potential expression of a negative bias, as expressed by Turkle when saying that simulated feelings are not real feelings (Turkle, 2010). This might in turn generate negative feelings in users who perceive their engagement as genuine leading them to scale back involvement and interactions with these agents. We propose that a similar approach could undermine the quality of social relationships with agents.

As previously mentioned, another objection refers to the 'ideal rationality' of intentional agents. What we perceive as intentional behavior in machines does not necessarily have a non-rational counterpart, as it is the case for biological intentional agents. Stich notes how intentional systems in Dennett's sense are unavoidably rational. In most cases, robots and virtual agents are not designed to have the option of irrational behavior and what is perceived as such is likely caused by errors or malfunctions. Hence, we argue that to come to terms with Stich's position, a solution must be found that allows people to treat agents as intentional when they behave rationally as long as this is beneficial for the interaction, and yet to switch to an alternative mechanistic approach if they observe (apparently) irrational behavior. This idea is supported by Wiese and colleagues, who call for the development of design strategies that support interactive flexibility. In other words, it must be possible to alternate between intentional and mechanistic mental models depending on one's specific interaction needs and contextual behavior (Wiese et al., 2017). Similarly, other authors argue that people treat robots alternatively as things or agents during different moments of an interaction (Alač, 2016).

We previously noted how adopting one or the other framework might not be always a conscious choice and that chances are that people interpret artificial agents' behavior socially as a default option. Therefore, the suggestion of developing design solutions that allow users to switch from one framework to the other should be considered under this assumption of a "primacy of the social mindset". In other words, while in certain cases it might be beneficial for the interaction to support the attribution of intentions and other mental states, other circumstances might require the opposite approach. In general, following Weick's sensemaking theory, it is important that the process of meanings co-construction (i.e., sensemaking) is lifted from the private and implicit sphere to a public and explicit level (Weick, 1995; Weick et al., 2005). This supports the notion of a more active involvement by users and the adoption of alternative semantics that support users' awareness, as we propose. This approach can also offer people a strategy to reduce the risk of wrongly adopting

one framework (e.g., the mentalistic one) instead of the other, which would lead to incorrect predictions when an artificial agent's behavior does not match the adopted mental model (Wiese et al., 2017). The next sections consider under what circumstances one or the other framework might be the most appropriate.

However, such considerations on ideal rationality do not necessarily imply that every time an agent's behavior leads to an unpredictable outcome from an intentional perspective, this is necessarily the result of a system error or malfunction. It might well be that the user simply cannot make sense of certain actions because she cannot immediately grasp the reasons behind them. It is in such cases that the user typically asks for an explanation (Miller, 2019). This can still be provided by the agent within an intentional framework, thus highlighting an "information asymmetry" according to which the agent's decision-making process was simply not obvious (Malle et al., 2007). Alternatively, the explanation could clarify whether an internal failure has occurred, thus letting users know that a mechanistic model would be more appropriate. In conclusion, this transition to non-intentional frameworks is likely to proceed more smoothly if users are made aware of the necessity to switch. We shall return and provide further support to this assumption in the last section where we discuss ethical implications.

2.4 AlphaGo: A Case Study

Here, we briefly discuss how the elements discussed so far are not only a matter of theoretical debate or experimental testing, but also apply to real life. To do so, we analyze a few aspects related to the case of Deep Mind's AlphaGo. One of the main reasons to reflect upon this case is that it offers an 'updated' direct comparison with Dennett's original example of the chess-playing computer. However, we can expect the future to provide further examples as this type of technology becomes more broadly present in our society.

Go is a very old board game and among the most complex ones, where 'human intuition' plays a fundamental role. Deep Mind's Go-playing system is not preprogrammed by expert players to perform a set of specific moves. Rather it is trained (or trains itself) through reinforcement learning. Through mimicking human strategies first, and then playing against different versions of itself (Silver et al., 2017), the system is able to improve and adapt its strategies autonomously. When challenged by some of the best human players, AlphaGo has repeatedly proved its efficacy in the game (Andras et al., 2018; Curran et al., 2019).

Curran and colleagues conducted a content analysis of how the Chinese and American press approached AlphaGo's games. Beside the predictable cultural differences, they also highlight how it is not unusual to attribute qualities such as 'intuition' and 'creativity' to the system (Curran et al., 2019). Furthermore, they argue, if such qualities "are no longer the sole domain of humans, there is a demand for a reconceptualization first and foremost of what it means to be human" (Curran et al., 2019, p. 733). In other words, they note, observing traits typical of human intelligence in a machine (whose nature is always transparent)

might even lead to an ontological reconsideration of what it means to be human (Severson & Carlson, 2010; Kahn, et al., 2011).

Another interesting aspect is the fact, that "some moves are made that are novel and inexplicable to human Go-playing experts and yet are effective, leading to more wins and new insights into the game" (Andras et al., 2018, p. 79). Few things can be noted. Heider differentiates intentional actions from unintentional events by saying that the former exhibit 'equifinality' (Heider, 1983). While AlphaGo can employ new and unpredictable moves, the apparent intention to win the game remains the same, i.e., oriented towards the same goal, i.e., 'equa-final' (Heider, 1983).

A second remark emerges that concerns and further explains the previously discussed distinction between perceiving and attributing intentions, particularly with those AlphaGo's moves that are inexplicable and yet effective (especially move 37 of the second match against Lee Sedol (Metz, 2016)). The reason why certain moves were difficult to predict is that human players would have hardly ever used them in those circumstances. The commentators of the game even wondered whether move 37 was a mistake. In pragmatic terms, what AlphaGo did was to opt for a very uncommon (among human players) strategy, whose outcome was a almost certain victory, although with a very small margin. Beyond uncovering new possible approaches to the game, the point we aim to make here concerns the fact that a move initially perceived as possibly erroneous turned out to be a winning one. In other words, the audience attributed the 'equifinality' of winning the game only in hindsight, while the initial perception was unclear. Furthermore, recalling the previous considerations on systems' ideal rationality, part of the audience was prone to attributing move 37 to a system mistake, highlighting the persistence of mechanistic interpretations of the system's behavior. However, such a possibility was later discarded as the move proved to be successful, although initially hard to predict and explain.

Finally, it can be noted that it would probably be very difficult for laypeople to obtain new insights into the game and learn new gaming strategies from a mathematical (i.e., design) perspective (for a similar analysis see Ling et al. (2019)). It might, however, be possible to interpret (or predict) those moves as if they had been made by a human whose goal is to win. In this regard, Curran and colleagues report a professional player commenting on one such move by saying that "almost no human would've thought of it" (Curran et al., 2019, p. 733).

In conclusion, Curran and colleagues observe how the type of narrative used by the media was influenced by the journalists' lack of domain-specific expertise, which allegedly led them "to relay on broad and undifferentiated frames" (Curran et al., 2019, p. 734). However, as we argued previously and partially in line with Dennett, a more plausible explanation is that the complexity of systems like AplhaGo does not leave laypeople (included most journalists and professional players) much room for interpreting the moves as the result of programming strategies. Rather, attributing to AlphaGo the intention (and desire) to win the match, the rationality needed to achieve this goal and the belief that a specific strategy would have been successful as it would be done with human players, is the only viable strategy for non-experts to understand, predict and perhaps learn from the system's behavior.

3 Tracing the Point of Origin of Intention Ascription in Artificial Agents: A Comparative Analysis of HCI and HRI

Dennett's original formulation of the intentional stance hinges on the complexity of sophisticated systems and the apparent rationality of their behavior as the main trigger for the ascription of intentions. Today's AI-based technologies are far more complex and advanced compared to those described by Dennett. Therefore, the issue of whether or not to treat today's machines as intentional agents is extremely pressing and relevant for their successful introduction into our society. How can complex rational behavior be unpacked and articulated to come up with implementable strategies? And how is this issue interpreted and studied in the empirical literature?

We approach this issue considering that attributing intentionality and other mental states to artificial agents is a flexible process (Abu-Akel et al., 2020). Furthermore, we acknowledge the intrinsic nuances of the process, which can very sensibly according to the already existing variety of artificial agents. According to Thellman and colleagues, how the attribution of intentions and other mental states varies depending on the type of agent represents an open challenge (Thellman et al., 2017). To this extent, they note, "there has been very little comparative research on how people actually interpret the behavior of different types of artificial agents" (Thellman et al., 2017, p. 1). Therefore, in this section, we analyze relevant examples from the experimental literature on virtual and embodied agents to investigate these assumptions from a comparative perspective. Based on this analysis, we argue that only by adopting a transversal approach does it become possible to grasp the nuances of this flexibility.

However, it is important to acknowledge that situations in real life might soon become even more nuanced, especially as the number of typologies and the diversification of artificial agents to interact with increase. Whereas we circumscribe the analysis to virtual and embodied agents, variants of each category already exist (e.g., anthropomorphic and machine-looking robots) that are worth examining individually and in comparison with other forms of social presence (Cassell, 2000; Bartneck, 2003; Kiesler et al., 2008; Li, 2015). Furthermore, there is in the literature a lack of long-term studies, which would help building a better understanding of how processes such as the attribution of intentions and other mental states evolve over time.

An initial distinction that emerges from the comparative analysis highlights how the phenomenon depends on intrinsic features of the agents, people's dispositions and external and contextual conditions. Furthermore, how the combination of these factors influences the overall process is rarely taken into consideration (Marchesi et al., 2019; Schellen & Wykowska, 2019).

3.1 Contextual Conditions

Several contextual elements that contribute to triggering the attribution of intentions can be identified. For instance, as illustrated by the example of AlphaGo, a society's cultural background can have repercussions for its perception of artificial agents

(Haring et al., 2014; Curran et al., 2019). Even more, attribution of anthropomorphic traits appears to be influenced by whether people perceive artificial agents to be members of the same in-group (e.g., in terms of nationality or gender) rather than of out-groups (Eyssel & Kuchenbrandt, 2012; Eyssel et al., 2012; Kuchenbrandt et al., 2013). Another relevant avenue of research investigates how attribution of anthropomorphic traits to physically present artificial agents can influence people's success in carrying out social and cognitive tasks (Riether et al., 2012; Spatola et al., 2019, 2019).

However, one element that occupies a central position is the type of tasks involved in the interaction (Epley et al., 2007; Marchesi et al., 2019). The fact that many artificial agents are meant to be employed in social contexts makes this aspect particularly relevant. For instance, Chaminade et al. (2012) conducted an fMRI study involving a competitive scenario (rock-paper-scissors) to compare attitudes towards the competitors—a human, an 'intelligent' robot and a 'random agent' (that did not base its moves on any strategy). Their results show that while participants treated the human competitor as being intentional, their reactions towards the robot were not significant in terms of intention attribution (Chaminade et al., 2012). As a possible explanation for this, the authors point to participants' lack of a clear cognitive strategy to interact with the agent, which resulted in them relying mostly (or exclusively) on individual opinions about the robot's inner mechanisms (Chaminade et al., 2012).

However, a different explanation is provided by Thellman et al. (2017), who suggest that the simple experimental scenario was the reason why no significant attribution of intentions was detected. While it is true that individual expectations do indeed play a central role (as discussed later), considering that a game like rock-paper-scissors does not involve much strategy (unlike other games, such as chess or Go), it becomes clear that the same type of agent might be treated as being either intentional or mechanical depending on the interaction affordances. This interpretation further refines the idea of people adopting a default social mindset when interacting with artificial agents. Specifically, the last consideration suggests that people tend to adopt a mentalistic approach as a default option when other cognitive processes are involved (e.g., strategic thinking and social cognition) (Spunt et al., 2015). However, the different interpretations of the results obtained by Chaminade et al. (2012) highlight another contextual factor we ought to consider.

Researchers' attitudes when investigating this phenomenon (or any phenomenon) might play a part in influencing how participants perceive an agent. This aspect seems to be largely underestimated in the literature. Perhaps this is because researchers' attitudes are not believed to have a direct impact on real social interactions. However, the way a researcher approaches a topic surely influences what can be found (i.e., when a researcher studies a phenomenon, the divergences and biases introduced by his or her unique point of observation may go unnoticed). Consequently, as researchers are among the people in charge of designing artificial agents, their approach findings can influence the societal perception of a specific topic in an indirect way, particularly in the longer term.

Another example stems from the analysis by Lim and Reeves (2010). The authors discuss levels of engagement in gaming experiences when playing with or against

'avatars' and 'agents'. Based on several studies, they state that when people believe they are interacting with a digital avatar of a real person, perceived social presence is higher compared to when interacting with an artificial agent. While in principle this might be a sound assumption, the authors hypothesize that a negative attitude towards the agents arises because players cannot 'mentalize' their opponent when this is an agent (rather than an avatar) (Lim & Reeves, 2010). In particular, their assumptions rest on a description of agents rooted in their (lack of) biological intentionality (Lim & Reeves, 2010). However, as we discussed previously, attributing intentions to an agent does not necessarily imply biological forms of intentionality. In conclusion, whereas relying on biological definitions of intentionality might explain negative dispositions towards agents, if engagement and ascription of intentions are detected and measured, this implies that people can and do mentalize artificial agents. Hence, the explanation provided by Lim and Reeves (2010) does not hold and, to the contrary, shows an underlying bias in addressing the topic.

3.2 Human Attitudes

We have previously noted how, alongside objective biological interpretations, intentionality (and the ability to infer and ascribe intentions to others' behavior) can also be read as a socio-cognitive construct (that makes social interactions possible). The intertwining of these two processual levels starts at very early stages of life. In fact, "when infants follow others in adopting the intentional stance, they acquire better interpretational resources, which increases their incorporation into the adult environment, and this, in turn, furthers the process of enculturalization." (Perez-Osorio & Wykowska, 2019). Furthermore, the ability to mentalize others might be part of a cerebral network labeled "the brain's deafult network", which has been shown to activate when one tries to mentally anticipate and explore social scenarios (Buckner et al., 2008). Consequently, people are trained to mentalize and recognize intentional patterns (Frith & Frith 1999, 2006; Chaminade et al., 2012; Perez-Osorio & Wykowska, 2019) and, more generally, to attribute anthropomorphic traits to non-human entitites (Caporael & Heyes, 1997; Nass & Moon, 2000; Nass et al., 1994), meaning that these strategies are widely available if necessary, according to the interaction affordances. We shall now discuss what it means for a strategy to be available if necessary.

Importantly, it can be noted that for people, it still makes a difference whether they interact with conspecifics or with artificial agents. In other words, brain activation is stronger in human–human interactions. However, reported differences can vary greatly from case to case (Thellman et al., 2017; Marchesi et al., 2019; Perez-Osorio & Wykowska, 2019). To this extent, an interesting perspective is provided by Bossi and colleagues, who conducted a study analyzing how brain activity in the 'resting state', i.e., when not engaged in a task, biases the perception of robots during interaction. They found that if mentalizing processes are present during the resting state, people are more likely to treat robots mechanistically later when interacting with them (Bossi et al., 2020). They explain these counterintuitive results by arguing that "if participants were involved in thinking about other people, and their

intentions or mental states in general, before they took part in the task, the contrast with a robotic agent might have been larger" (Bossi et al., 2020, p. 4). Hence, although the attribution of intentions is a strategy that is always available, its adoption (or lack thereof) might be affected by the preceding neural activity, showing a non-linear correlation with other variables such as the type of activity or the general disposition of individuals towards artificial agents.

Despite quantitative differences, there seems to be a certain degree of agreement on the possible cause for this differential activation of mentalistic schemata. The idea is that it is fairly easy for people to interpret certain artifacts as material objects and humans as intentional agents. Everything in between lacks a specific ontological categorization, forcing people to adopt a familiar framework, which often turns out to be the intentional one (Davidson, 1999; Thellman et al., 2017; Marchesi et al., 2019; Abu-Akel et al., 2020). Consequently, as previously argued, people adopt this strategy when it proves to be the most efficient or reliable (Perez-Osorio & Wykowska, 2019). To this extent, according to Weick, if previously adopted sense-making strategies have been successful, they will be retained and reenacted in future interactions (Weick, 1995). However, this is likely to not always be the most fruitful approach, but only in cases where tasks require social cognition (Epley et al., 2007; Spunt et al., 2015; Wiese et al., 2018; Ohmoto et al., 2018; Schellen & Wykowska, 2019). It is in such cases that treating artificial agents as intentional tends to improve the quality of the interaction (Wiese et al., 2017; Schellen & Wykowska, 2019).

Furthermore, in line with Dennett's idea of complexity, it should also be considered that approaching artificial agents from a mechanistic perspective is generally difficult for many people (Dennett, 1981). The reason is that it may be cognitively too demanding, especially for non-expert users, to try to make sense of artificial agents behavior from a mathematics-based, design stance. This highlights a connection between the concept of systems' complexity and the idea of necessity. However, such a relationship should be read in light of the type of tasks and interactions involved, as previously discussed. Before drawing any conclusion, the next paragraph will consider the last set of relevant elements that can influence the attribution of intentions and other mental states. Additionally, it should still be considered that cultural or personal dispositions towards technology might still override the availability of an intentional framework, encouraging people to adopt either a mechanistic or an anthropomorphic approach (Waytz et al., 2010; Haring et al., 2014).

3.3 Intrinsic Features

Analyses of internal or exhibited features of agents that factor into the attribution of intentions (and more broadly of social skills) converge towards two categories of qualities: appearance and behavior (Wiese et al., 2017). This, supported by advances in neurosciences, particularly the availability of fMRI techniques, has led some researchers, especially in the field of human–robot interaction, to emphasize the importance of anthropomorphic embodiment as a trigger for the use of mentalistic descriptions (Marchesi et al., 2019). Of particular interest, in this direction is the discovery of mirror neurons (Rizzolatti & Craighero, 2004). These appear to play a

role in the processes of attributing intentionality based on embodiment, so that similarity to human physical presence triggers higher activation.

Consequently, research on physical appearance has focused on endowing agents (particularly robots) with human-like features. Some features, such as a face (Johnson, 2003; Looser & Wheatley, 2010; Balas & Tonsager, 2014), gazing eyes (Khalid et al., 2016; Willemse et al., 2018), a non-symmetric ratio between the head and the body, smooth bodily transformations as opposed to rigid and linear changes, (Johnson, 2003), and the visibility of the entire body (Chaminade & Cheng, 2009) have been highlighted as preeminent. This approach is rooted in the idea that "humans might be able to understand the behavior of human-like robots more easily than, for example, the behavior of autonomous lawnmowers or automated vehicles" (Thellman et al., 2017, p. 2). This is a plausible explanation, as feature similarity is likely to more effectively and quickly activate the brain areas involved in mentalizing and motor resonance (Chaminade et al., 2007; Wiese et al., 2017).

However, if the attribution of intentions mostly depends of appearance, this would not explain positive results in the absence of a body or with very different forms of embodiment. Interestingly, Ziemke (2020) reports one such case in relation to a road accident involving an autonomous vehicle. The accompanying report by the U.S. National Transportation Safety Board notes how some people were surprised that "Uber's self-driving car didn't know pedestrians could jaywalk" (Ziemke, 2020, p. 1). This is explained as "an expectation-probably shared by many people that driverless cars should have a human-like common sense understanding of human behavior." (Ziemke, 2020, p. 2) More generally, this implies that behavioral elements might be at least as important as appearance, if not more (Terada et al., 2007; Wiese et al., 2017). Among qualities in this category, studies have focused on the reciprocity and contingency of the behavior in relation to the environment and to other agents (Johnson, 2003; Pantelis et al., 2014, 2016). Furthermore, autonomous, rational and seemingly biological motion play a central role (Castelli et al., 2000; Gazzola et al., 2007; Oberman et al., 2007; Pantelis et al., 2016; Abu-Akel et al., 2020).

In this respect, experiments in the tradition of a well-known study by Heider and Simmel make an important contribution. Heider and Simmel argued how people attribute social skills to geometric shapes in motion (Heider & Simmel, 1944). Developing this concept further, Pantelis and colleagues analyzed the relationship between the goal-directed motion of similarly simple, autonomous geometric objects in a two-dimensional virtual environments and the attribution of mental states (Pantelis et al., 2014, 2016). The results of the first study show that people tend to estimate agents' states (e.g., when they are 'attacking' another agent, or 'fleeing' from it) correctly and coherently with one another (Pantelis et al., 2014). Perhaps more revealing are the results of a follow-up study, where an evolutionary factor is introduced into the agents' behavior. In fact, the authors hypothesize that the ascription of mental properties is at least partially related to how artificial agents adapt to their environment (Pantelis et al., 2016). One of their main arguments is that people's ability to correctly infer the agents' states increases concurrently with the rationality of the agents' behavior. Their results show that people tend to infer more accurately the mental states of agents that adapt their behavior. These studies not only

corroborate the relevance of motion for appropriate judgements of the behavioral information that motion itself conveys. Arguably, they support the idea of a primacy of 'pro-social rational content' over the means of conveyance (i.e., in this case, motion itself). Thus, in the absence of such minimal rationality, artificial agents' behavior does not communicate any mental state (Pantelis et al., 2016).

This last proposition is supported by the fact that appearance and biological motion alone (or combined) cannot explain the ascription of intentions and the activation of brain areas responsible for mentalizing in cases where both features are missing. The study conducted by Abu-Akel et al. (2020) is in line with this position. The authors investigate the ascription of intentions to a virtual agent in a competitive scenario, with the participants not able to see their competitor. They hypothesize that the activation of brain areas involved in mentalizing operations does not require motion. Instead, they claim that abstract information about the opponent is sufficient as long as it is considered an intentional and rational agent of either natural or artificial nature. Thus, their results show how "activation of the 'mentalizing network' might be specific to mentalizing, but it is not specific to mentalizing about humans." (Abu-Akel et al., 2020, p. 8). Interestingly, conclude the authors, "such flexibility in the attribution of intentionality (whether to active or passive, human or computer agents) can be manipulated volitionally and even strategically" (Abu-Akel et al., 2020, p. 8). As it will be addressed in the next section, this has have ethical implications.

Another study pointing in a similar direction was conducted by Pinchbeck (2008). Here, the authors analyze how to enhance gaming experiences by implementing simple behavioral tricks in non-player characters, rather than relying on more complex AI techniques to drive more nuanced individual behaviors. Referring to a group of non-player characters (human mercenaries), they describe a 'breakdown of intentionality' as a consequence of the characters' incoherent behavior. Under certain conditions, these characters enter a 'combat state' (i.e., 'seemingly intentional' behavior of actively seeking enemies). Instead, when they are in the water they engage in a sort of 'rest state' ('the pool party effect', in the terminology used in the paper), making themselves vulnerable to attacks. Furthermore, the characters seem to be uninterested in solving this issue, an attitude that the authors identify as totally irrational. This negatively affects the attribution of intentions and rationality to the characters (and the gaming experience). This, the authors note, happens because people tend to grant intentionality when actions are recognized as ecologically valid (Pinchbeck, 2008). By contrast, another kind of non-player character (i.e., mutated monkeys called Trigens) display more ecologically valid behavior by avoiding entering the water, which would cause them to drown. This rational attitude appears to suggest a higher degree of intentionality despite the characters' less human appearance.

In conclusion, as the examples show, our analysis identifies a few concepts that are useful for the design of artificial agents. Implementing features that support the attribution of intentions can be a desirable strategy, as it may enhance the overall quality of the interactions. In this way, manifest behavior that conveys a message of contextual, pro-social rationality serves as the main spark that ignites the processes of attributing mentalistic qualities to artificial agents. This is supported by the fact

that the ascription of intentions and other mental states is a widely available mental process that people are trained to engage in beginning at an early age. Hence, this also clarifies our previous point on necessity. While the ontological classification of most objects is not problematic, as soon as more sophisticated devices' behavior appears as minimally rational and pro-social, the combination of the availability of a mentalistic approach and the likelihood of a cognitive overload that might derive from trying to make sense of such machines from a mechanistic perspective result in the default adoption of mentalistic schemata.

Whenever they can be implemented, features like a human-like appearance or biological motion are fundamental tools to support the process, and as such, they should always be considered as a possible design strategy. Nevertheless, human-like appearance and motion alone are not sufficient conditions (e.g., a highly anthropomorphic robot that does not act in a contextually rational way is likely to be treated as a sophisticated mannequin). They need to be accompanied by (appearance) or convey (motion) some sort of rational message with ecological validity. Accordingly, artificial agents that do not display either of these qualities (appearance or motion) can still be treated as intentional, as is for instance the case with AlphaGo or conversational agents. As the perceivable rationality of agents' behavior increases, the attribution of intentions becomes more likely. Additionally, as the interactions with artificial agents increase in number and variety, the attribution of intentions and other mental states may become part of implicit social cognition processes. Referring to models of the mind proposed within the context of "dual processes" and "dual systems" theories, this would further reduce the cognitive load, and make the process more automatic (Evans & Stanovich, 2013).

However, it is important to note that depending on the interaction context, the tasks to be carried out and the type of machine, a mentalistic approach might not always be the most appropriate. If no social cognition is involved (e.g., as with autonomous vacuum cleaners or lawnmowers), more mechanistic mental models are adequate. Referring to design features that allow users to switch from one interpretative framework to another, it is fundamental that such design strategies consider said transition in both directions. For instance, if a robotic vacuum cleaner crashes or malfunctions, adopting a mentalistic approach is counterproductive. More generally, even when such machines function properly treating them as intentional agents would likely not be very beneficial to the interaction. This further supports the emphasis on social cognition and the idea of seemingly rational behavior as triggers. Similar considerations are to be accounted for in the design phase of artificial agents, also in light of the fact that people tend to attribute human traits to machines even when, in principle, mechanistic approaches would be more appropriate (Carpenter, 2013).

4 Ethical Considerations: Attribution of Intentions and Deception

This section of the paper takes a twofold approach to the ethical aspects of adopting the intentional stance. On the one hand, we acknowledge that, as artificial agents' presence in a growing number of everyday contexts increases, it is important that

interactions with them become progressively more efficient, pleasant and trustworthy. Accordingly, design strategies that support users' adoption of the intentional stance in contexts that involve social cognition might be sought after (Spunt et al., 2015; Schellen & Wykowska, 2019), particularly with respect to mutual connections, joint human–agent efforts, and, more generally, social acceptance of artificial agents (Wiese et al., 2017). On the other hand, these efforts aimed at improving the overall quality of human–agent interactions should not translate into design strategies that make the categorization of artificial agents ambiguous (Hackel et al., 2014; Mandell et al., 2017). In fact, extreme anthropomorphic attributions might have a negative impact on the quality of the interaction (Mandell et al., 2017; Ziemke, 2020) if, for instance, people start perceiving artificial agents as a threat rather than valuable resources (Spatola & Normand, 2020). It is therefore important to find a proper balance between these two necessities in advance, so as to not leave the burden of evaluation entirely on the people interacting with the agents.

4.1 Layers of Deception in Human–Agent Interaction

Before analyzing whether the implementation of features that resemble intentional behavior should be labeled a deceptive design strategy, it is first necessary to briefly consider what deception means in the first place. According to Danaher, at the lowest level of analytical granularity, "deception involves the use of signals or representations to convey a misleading or false impression. Usually the deception serves some purpose other than truth, one that is typically to the advantage of the deceiver." (Danaher, 2020, p. 118) According to this interpretation, deception centers around three main elements: the person being deceived, the agent directly responsible for perpetrating the deception, and the signal or misleading information. Another layer must be considered. It is represented by the interests of what we call a 'third party', which typically is the entity or set of entities (e.g., companies, designers, malicious users etc.) that act from 'behind the curtain' to provide the conditions necessary for the agent to perform deceptive acts. These are often the actors that ultimately gain the greatest advantage, for instance, in terms of data use (Kaminski et al., 2016; Hartzog, 2016) or for unethical commercial or even criminal purposes (O'Leary, 2019). It is important to acknowledge this aspect within the context of human–agent interaction, for reasons that are mostly related to responsibility distribution, as it will be addressed further on.

The issues with attributing intentions to artificial agents and the implementation of strategies meant to trigger such attributions are at the heart of the debate on deception. The first consideration in this regard is quite nuanced. As Danaher notes, it has to do with the fact that what exactly constitutes a deceptive act among humans is defined by the intentions, desires and beliefs of the deceiver. Arguably, taking such a perspective in human–agent interaction is problematic, since whether agents have intentions and other mental states or not is itself part of the debate on deception (Danaher, 2020). Most of the debate concentrates on interpretations of intentionality that are in line with or similar to Searle's. Often, this type of criticism is also directed at features that express emotional engagement (such as care

or love). Therefore, anthropomorphic cues that do not reflect actual qualities (e.g., mental states) are fundamentally seen as deceptive. Some authors argue that anthropomorphic behavior and 'simulated qualities' are designed to trick and fool people precisely because they let people believe robots (and other agents) have those qualities that they lack (Sparrow & Sparrow, 2006; Sharkey & Sharkey, 2010; Turkle, 2010; Elder, 2016). More broadly, it is argued that the implementation of features that express seemingly intentional behavior could trigger categorical uncertainty (in ontological terms) and therefore undermine social interaction (Hackel et al., 2014; Mandell et al., 2017).

Not only researchers but also institutions that oversee the development of AI and robotics have highlighted the potentially negative aspects of excessive anthropomorphization/personification of artificial agents. The EU High Level Group's call for trustworthy AI is one such example, but several other bodies have moved in a similar direction (Coeckelbergh, 2019; Floridi, 2019; HLEG, 2021). For instance, The UK Engineering and Physical Sciences Research Council's (EPSRC) Principles of Robotics clarifies that robots should not be designed to deceive users and should always be clear and transparent about their artificial nature (Theodorou et al., 2016; Boden et al., 2017). Similar efforts highlight that robots and other artificial agents should not pose as humans (Shahriari & Shahriari, 2017; Heaven, 2018).

However, it is important to note that not all researchers agree with these positions. In fact, some consider at least some forms of deception to be an intrinsic feature of robotics and AI, as they offer the best possibility of successfully developing socially integrated artificial agents. As such, deception is seen as an acceptable, even desirable phenomenon (Wagner & Arkin, 2011; Shim & Arkin, 2012; Isaac & Bridewell, 2017; de Oliveira et al., 2020). Indeed, one might argue that deception even lies at the foundation of the Turing test, the many versions of which share the assumption that, in order to pass the test, a machine must succeed in convincing a human jury that they are actually interacting with another human.

Danaher highlights a further possible distinction. He regards what he calls 'hidden state deception' as the most dangerous layer. This form of deception occurs when agents hide capacities they possess by means of deceptive signals (Danaher, 2020). Collecting personal data without users knowing it or, even worse, pretending it is not happening falls into this category. While it is reasonable to share such concerns, this paper primarily aims to discuss another level of potential deception, what Danaher calls 'superficial state deception'. This entails that an agent "uses a deceptive signal to suggest that it has some capacity or internal state that it actually lacks." (Danaher, 2020, p. 121) Indeed, implementing features that resemble intentionality is a form of superficial state deception, although the main beneficiaries of the two levels are roughly the same (i.e., the aforementioned third parties).

4.2 Seemingly Intentional Behavior is not (Necessarily) Deception

The thesis we defend here is a twofold one. First, we argue that in principle, it is not unethical to opt for design strategies that support the adoption of the intentional stance. However, to avoid feelings of deception, a fundamental prerequisite is that

users are put in a position to, if not consciously decide, at least be aware of the nature of the agent. We approach this discussion based on the aspects introduced in the first part of this section and in light of the alternative approach previously proposed. Regarding the recursive argument pinpointed by Danaher, we identify two levels of interpretation. If to say that someone is a deceiver implies the intention to deceive, then we argue that this is not the case with artificial agents. By speaking of 'seemingly intentional behavior', we mean to emphasize the conscious attempt to emulate human behavioral traits for the sake of interactional quality. As such, the term specifically seeks to avoid conflation with biological, evolution-based interpretations of intentionality. Artificial agents do not have genuine intentions (e.g., to deceive) in the biological sense, which implies that they cannot be genuinely deceptive.

However, we also introduced the idea of a possible involvement of third parties. Artificial agents can, in principle, stand for the interests of said actors. Accordingly, in addition to what was previously reported, Jacob states that if "a speaker utters words from some natural language or draws pictures or symbols from a formal language for the purpose of conveying to others the contents of her mental states, these artifacts used by a speaker too have contents or intentionality." (Jacob, 2019, p. 1) We consider it problematic if these actors seek to deceive users through the agents (i.e., the artifacts that convey the actor's intention). As such, artificial agents could well be 'tools of deception' by these third parties.

For this reason, we deem it fundamental to make an ethical distinction between the promotion of illusions such as personification and the implementation of features that trigger the ascription of intentions. Recalling that intentionality is not only an objective quality, but also a social construct that makes interactions possible and even increases the overall quality of the relationships, we cannot consider the implementation of features that aim to resemble intentionality in itself as ethically problematic. Accordingly, the attribution of intentions should be seen as a strategy that people can adopt to better predict and interpret agents' behavior and to navigate social interactions with them. From this perspective, design strategies that facilitate this process should be regarded as worth striving for. Furthermore, one may even argue that if users feel more at ease treating agents as intentional in certain cases (as the experimental literature shows), telling them that this feeling is part of a deception could negatively affect their social interactions with the agents. In other words, similarly to what we argue about terms such as 'simulated intentionality', 'ethically exclusivist' positions could dissuade users from engaging in meaningful interactions with the agents if the only interpretative and predictive framework they can use or that seems to work is the intentional one. However, whether artificial agents should employ emotional terminology (such as 'I care'), for instance, is open to debate. In fact, such solutions may lead to the perception of ontological ambiguity (Hackel et al., 2014; Mandell et al., 2017).

4.3 Users' Awareness

The other point we highlight is the concerning possibility of artificial agents acting deceptively on behalf of third parties. As previously mentioned, we consider it

among the most pressing aspects of the present debate. Therefore, here we address the idea of 'promoting the illusion of personification'. In accordance with regulatory institutions that have called for greater transparency, we argue that users' awareness should be a *conditio sine qua non* for the design phase. In other words, for the implementation of specific features to be ethical and successful in fostering social engagement, people should be made aware of the nature of the agents they interact with. We also consider this a prerequisite for people to be given the possibility to switch to mechanistic approaches when necessary. Furthermore, we claim that building such awareness is the specific responsibility of the third parties in charge of the design of artificial agents. Coeckelbergh (2018) takes a relevant position with respect to this issue stating that although deception can be seen as a co-created performance, designers and other third parties have the responsibility to ultimately reveal the 'trick' behind it. In other words, they are responsible for the performative affordances of the agents they introduce into society (Coeckelbergh, 2018).

In accordance with Coeckelbergh, we consider performance as a co-construction (Coeckelbergh, 2018). An agent behaves in a seemingly intentional manner; a person then attributes intentionality to the performed act, and together, the two co-construct the interaction. However, we further claim that the tricks performed by an artificial agent should be revealed beforehand. It makes it possible to still be meaningfully engaged without necessarily thinking that the agent has intentions (in the biological sense) or, even worse, being surprised that one is interacting with an artificial agent. This last point is fundamental for a simple reason. The more sophisticated artificial agents become, the more difficult it will be for people to tell the difference (between a human and a machine) in advance. In order to shed greater light on our position, we now provide an example concerning the well-known topic of the 'uncanny valley'.

4.3.1 The 'Uncanny Valley' Case

The 'uncanny valley' hypothesis posits that extreme anthropomorphism can trigger negative reactions in people (Mori et al., 2012). In this view, the 'valley' of the curve represents the point where human-like appearance and behavior do not quite reach total resemblance, but are still enough to trigger rejection mechanisms. Many hypotheses have been proposed to explain the phenomenon. One argument in line with our position centers around the idea of rejection as the expression of violated expectations (Saygin et al., 2012; Urgen et al., 2018). Urgen and colleagues conducted a study with a highly anthropomorphic android. Whereas the android was human-like enough to generate high initial expectations through its appearance alone (e.g., by looking at a picture or a video of it), as soon as the android began to move its artificial nature became evident, triggering feelings of rejection. This connection between movement and eeriness or uncanniness was previously raised by Mori et al. (2012), but not investigated thoroughly.

Urgen and colleagues argue that, although participants generate an initial mental model (and expectations) based on appearance alone, when the robot's movements reveal its true nature, the established model fails to hold, contributing (in this case) to an increase in uncanny feelings (Urgen et al., 2018). A similar mismatch between

appearance and motion is reported in Saygin et al. (2012). Also in this case, participants were exposed to videos of different types of agents, while their neural activity was monitored. Respectively, the agents were a robot with mechanical appearance, an android and a human. They conclude that while "the android used in our study is often mistaken for a human at first sight, longer exposure and dynamic viewing has been linked to the uncanny valley" in reason of the participants' prediction error that the mismatch generates (Saygin et al., 2012, p. 420). Interestingly, while in Saygin et al. (2012), Urgen et al. (2018), participants are shown only videos of the robot, in other studies with similar androids people interact directly with them (Bartneck et al., 2009). Importantly, in this last case people are always aware of the fact that, regardless of how extremely anthropomorphic, the agent is artificial. In the previously cited cases, they are not. They only become aware of this when the robot moves.

The different records in terms of feelings of uncanniness reported by the studies can be explained by the abrupt failure of expectations (reported in the first two studies). When the conditions for behaving 'as if' are not made explicit, people (in the considered case) are likely to simply behave as they would with other humans. But when the illusion is broken, so are the mental models, generating feelings of rejection.

4.4 'Third Parties" Responsibility

These considerations support our claim that users need to be made aware of the nature of the agent and that this awareness contributes to the quality of the interaction. Thus, our final point for reflection concerns our last claim, that responsibility for ensuring users' awareness of the nature of an artificial agent should fall on the 'third parties' in charge of designing what kind of performance the agent is capable of providing. This issue, notes Coeckelbergh, is seldomly considered in robotics, because "the designer (and especially the company) needs to sell the device as magic" (Coeckelbergh, 2018, p. 80). Fundamentally, in order to have the chance to switch to behaving 'as if', people must be aware of the type of performance that has been created. At the same time, treating agents 'as if' does not mean that there is nothing real to be gained from interacting with them. The thoughts, impressions and feelings one experiences are real, rather than 'simulated' (Turkle, 2010; Seibt, 2017).

Furthermore, not only should ensuring such awareness be third parties' responsibility but, as we noted, the underlying trick should be revealed before the interaction takes place. In fact, as artificial agents become progressively more sophisticated, the decision to automatically behave 'as if' might become less obvious. This is particularly the case with virtual technologies like conversational agents and chatbots. In fact, in most cases it is still possible and fairly easy to determine the artificial nature of a physical robot. No matter how well crafted modern androids might be, it is particularly difficult to flawlessly replicate an extremely human-like appearance, the smoothness of human movements, non-verbal cues, etc., so that, as in the considered example, only through an indirect medium (i.e., pictures or videos) and in

absence of movement could said androids be mistaken for humans. The same cannot be taken for granted when the interaction occurs in a fully virtual environment, as the next examples show.

4.4.1 Google Duplex and GPT-3

Google Duplex is a conversational agent endowed with natural language features to handle tasks such as making reservations and appointments (O'Leary, 2019). The most relevant aspect here is that the system does not only engage in natural language conversations. It also incorporates what (O'Leary, 2019) calls 'speech disfluencies', conversational elements that break the flawless pace of a conversation (such as 'hmm's' and 'uh's'). These kinds of 'interruptions' are very common in human–human interactions, because people do it often as they try to gather their thoughts (Leviathan & Matias, 2018). Nevertheless, including such disfluencies has attracted criticism, precisely because such behavior can be interpreted as deceptive. Consequently, Google's design choices have generated ethical concerns (Lomas, 2018). However, as we have already argued, conducting conversations by employing fluent natural language capabilities or even displaying sophisticated 'speech disfluencies' are not themselves the problem. As previously noted, such features could be worth striving for, as they can improve the overall quality of interactions. However, what is being debated here is the lack of a specific form of transparency that supports users' awareness of the kind of agent they are interacting with. Therefore, having the agent (Duplex, in this case) identify itself at the beginning of an interaction seems to be a reasonable solution (Bay, 2018; O'Leary, 2019). It should then be up to users to decide whether they want to continue under the specified conditions, i.e., once they have been put in the condition to behave 'as if'.

Another technology holding similar potential is GPT-3 (Generative Pre-trained Transformer 3) (Damassino & Novelli, 2020). This deep learning-based natural language processing model can generate text that is often indistinguishable from something a human would write. In their commentary, Floridi and Chiratti show the possibilities and limits of this tool (Floridi & Chiriatti, 2020). They note that in many cases, people might not recognize or even care whether a piece of text has been written by an artificial agent. While this might certainly be true, at least in the very near future, we believe that regulations should not only apply to ideal cases, but also to extraordinary ones. In other words, regulations requiring such artificial agents to make their nature clear in advance are likely to be necessary, especially in cases that could create ambiguity. Consequently, we agree with (Floridi & Chiriatti, 2020)'s conclusions that people should be able (i.e., put in the conditions) to discern what is what. One early example of the dual nature of this point is the use of GPT-3-powered bots on Reddit, a popular online platform. In fact, one of these bots was active under a normal username with almost no one noticing it. While most of its comments were reportedly unharmful, the bot also engaged in conversations about sensitive topics, such as suicide (Heaven, 2020). The bot's real nature was discovered when its text outputs were compared to those of the so-called 'Philosopher AI', another GPT-3 based bot (Heaven, 2020). The main difference between the two

is that in the case of the 'Philosopher AI' the artificial nature of the bot has been made clear from the beginning, allowing users to engage in entertaining question and answer sessions with the bot (including about the nature and the coding of the bot itself).

As a closing remark, a relevant concept for our last claim is represented by the process of dehumanization, as opposed to that of anthropomorphization. Dehumanizing is typically intended as failing to attribute human-like traits to other humans (Haslam, 2006; Epley et al., 2007; Haslam & Loughnan, 2014). Hence, in human–human interaction it often comes with negative connotations. However, in the context analyzed here (i.e., human–agent interaction), the idea of dehumanizing agents has positive consequences, at least as long as it is meant to counter the default anthropomorphizing trend. To this extent, Haslam and Bain note how a concrete (rather than abstract) mental 'construal' of other people could help reducing such a dehumanization process (Haslam & Bain, 2007; Haslam & Loughnan, 2014). Additionally, we previously noted how the lack of an agent-specific ontological categorization is among the main causes that trigger the attribution of anthropomorphic traits. Therefore, we assume that, in specific circumstances, supporting a more concrete mental 'construal' of artificial agents will let people engage in a specular process of dehumanizing artificial agents. To this extent, we claim that specific design features, such as endowing the agents with machine-like traits, having them identifying themselves as artificial entities, or pointing out the 'mechanical' nature of malfunctions, would support this process of dehumanization. In turn, this will help the switch from a mentalistic approach to a mechanistic one, when the latter is more pragmatically or ethically appropriate.

Hence, whereas some researchers' concerns about what is unethical for artificial agents may be too extreme, this last reflection leads us to agree with regulatory attempts to require companies and other third parties to adopt an approach that makes people aware of the type of performances displayed and the artificial nature of the performer. In principle, this paper calls for pre-performance forms of transparency, i.e., the nature of the agents' 'tricks' should be revealed before they are performed. This is at least partially in line with those researchers and institutions that promote higher transparency and, consequently, a distribution of responsibility that calls for explicit commitments by third parties. However, the important role played by machine seemingly intentional behavior in enhancing the quality of social interactions must be acknowledged. Therefore, we conclude that, as long as users are made aware of the nature of the agent they are interacting with, the implementation of strategies that support the attribution of intentions to those artificial agents meant to be employed in contexts that involve social cognition and skills should be considered not only ethically acceptable, but also ethically desirable. On the other hand, mentalistic frameworks appear to be the default approach that people resort to while interacting with seemingly rational artificial agents that do not clearly fall into objectual ontological categorizations. When no social cognition is involved, an opposite dehumanizing approach is more adequate. This could be pursued by, for instance, emphasizing the artificial nature of the agents and their machine-like traits.

5 Conclusions and Limitations

The main aim of this paper is to discuss why the attribution of intentions is effective and desirable and to identify corresponding design suggestions. To do so, we propose an analysis in three main directions, corresponding to the three main sections of the paper. First, we discussed semantic implications of the concept, in light of definitions of intentionality and of some objections directed at Dennett's idea, with particular attention to the behaviorism that informs it. We emphasized how the notion that the intentional stance is a strategy to understand and predict the behavior of sophisticated artificial agents represents the most useful aspect of Dennett's formulation. This led us to suggest the adoption of alternative terminology, in order to reduce the risk of conflation between the attribution of intentions to artificial agents and biological approaches to intentionality.

Furthermore, we traced the point of origin of the process of intention attribution by examining experimental literature about robots and virtual agents. Our conclusion is that contextually valid rationality represents the most important feature in order for agents to be treated as intentional. However, we also identified how this can and should be supported by other features and contextual conditions. Additionally, we considered how a mentalistic approach is not the most appropriate when no social cognition is involved and suggested possible strategies to counter excessive anthropomorphization accordingly.

Finally, we discussed possible ethical implications of the attribution of intentionality to artificial agents. While acknowledging the possible benefits of an intentional framework for social engagement with agents, we also identified a prerequisite for the ethical acceptability of such a framework. In line with most regulatory institutions, we argue that is necessary for users to be made aware of the agents' artificial nature and provide examples to support our claim.

One question that we leave open concerns to what extent the actual implementation of features that trigger the ascription of intentions and other mental states should be pushed. Referring to the case of Google Duplex, we said that the speech disfluencies employed by the system are not themselves the problem. Is the same true for the use of 'more sensitive' and openly mentalistic terms like 'I understand', 'I think', or their emotional counterparts like 'care' 'love', etc.? The risk we identify in this case is an extreme and perhaps 'deceptive' form of anthropomorphism. Where should the line be drawn? We question the necessity to employ such emotionally and semantically rich terminology for the ascription of intentions and other mental states to be successful. Perhaps, a critical analysis of such issues in light of other, related concepts, such as that of a "phenomenal stance" will help shed greater light on the debate.

Funding Open access funding provided by TU Wien (TUW).

References

Abu-Akel, A. M., Apperly, I. A., Wood, S. J., & Hansen, P. C. (2020). Re-imaging the intentional stance. *Proceedings of the Royal Society B, 287*(1925), 20200244.

Alač, M. (2016). Social robots: Things or agents? *AI & Society, 31*(4), 519–535.

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., & Milanovic, K. (2018). Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine, 37*(4), 76–83.

Balas, B., & Tonsager, C. (2014). Face animacy is not all in the eyes: Evidence from contrast chimeras. *Perception, 43*(5), 355–367.

Bartneck, C. (2003). Interacting with an embodied emotional character. In *Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces* (pp. 55–60).

Bartneck, C., Kanda, T., Ishiguro, H. & Hagita, N. (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* (pp. 269–276). tex.organization: IEEE.

Bay, M. (2018). *Am I speaking to a human?*, Retrieved May 10, 2018, from https://slate.com/technology/2018/05/google-duplex-can-make-phone-calls-for-you-but-it-should-have-to-identify-itself (tex.jorunal:slate).

Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review, 90*(1), 5–43.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., & Kember, S. (2017). Principles of robotics: Regulating robots in the real world. *Connection Science, 29*(2), 124–129.

Bossi, F., Willemse, C., Cavazza, J., Marchesi, S., Murino, V., & Wykowska, A. (2020). The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. *Science Robotics, 5,* 46.

Breazeal, C. L. (2002). *Designing sociable robots*. MIT Press.

Breazeal, C., & Scassellati, B. (1999). How to build robots that make friends and influence people. In *Proceedings 1999 IEEE/RSJ international conference on intelligent robots and systems. Human and environment friendly robots with high intelligence and emotional quotients (cat. No. 99CH36289)* (Vo.l 2, pp. 858–863). (tex.organization: IEEE).

Buckner, R., Andrews-Hanna, J., & Schacter, D. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences, 1124,* 1–38.

Caporael, L. R. (1986). Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior, 2*(3), 215–234.

Caporael, L. R., & Heyes, C. M. (1997). Why anthropomorphize? Folk psychology and other stories. Anthropomorphism, anecdotes, and animals, 59. State University of New York Press

Carpenter, J. (2013). The Quiet Professional: An investigation of US military Explosive Ordnance Disposal personnel interactions with everyday field robots (Unpublished doctoral dissertation).

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM, 43*(4), 70–78.

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage, 12*(3), 314–325.

Chaminade, T., & Cheng, G. (2009). Social cognitive neuroscience and humanoid robotics. *Journal of Physiology-Paris, 103*(3–5), 286–295.

Chaminade, T., Hodgins, J., & Kawato, M. (2007). Anthropomorphism influences perception of computer-animated characters' actions. *Social Cognitive and Affective Neuroscience, 2*(3), 206–216.

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutscher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience, 6,* 103.

Coeckelbergh, M. (2018). How to describe and evaluate "deception'' phenomena: Recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn. *Ethics and Information Technology, 20*(2), 71–85.
Coeckelbergh, M. (2019). Artificial intelligence: Some ethical issues and regulatory challenges. *Technology and Regulation,* 31–34.
Curran, N. M., Sun, J., & Hong, J. W. (2019). Anthropomorphizing AlphaGo: A content analysis of the framing of Google DeepMind's AlphaGo in the Chinese and American press. AI & Society, 1–9. Springer
Damassino, N., & Novelli, N. (2020). *Rethinking, reworking and revolutionising the turing test*. Springer.
Danaher, J. (2020). Robot Betrayal: A guide to the ethics of robotic deception. Ethics and Information Technology, 1–12. Springer.
Davidson, D. (1999). The emergence of thought. *Erkenntnis, 51*(1), 511–521.
De Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.
de Oliveira, E., Donadoni, L., Boriero, S., & Bonarini, A. (2020). Deceptive actions to improve the attribution of rationality to playing robotic agents. *International Journal of Social Robotics,* 1–15. Springer.
Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy, 68*(4), 87–106.
Dennett, D. C. (1981). *Brainstorms: Philosophical essays on mind and body*. MIT Press.
Dennett, D. C. (1988). Précis of the intentional stance. *Behavioral and Brain Sciences, 11*(3), 495–505.
Dennett, D. C. (1989). *The intentional stance*. MIT Press.
Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy, 88*(1), 27–51.
Dennett, D. C. (1993). *Consciousness explained*. Penguin.
Dennett, D. C. (1995). The unimagined preposterousness of zombies.
Dennett, D. C. (1997). True, believers: The intentional strategy and why it works. *Mind Design,* 57–79.
Dreyfus, H., Dreyfus, S. E., & Athanasiou, T. (2000). *Mind over machine*. Simon and Schuster.
Elder, A. (2016). False friends and false coinage: A tool for navigating the ethics of sociable robots. *ACM SIGCAS Computers and Society, 45*(3), 248–254.
Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864.
Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.
Eyssel, F., De Ruiter, L., Kuchenbrandt, D., Bobinger, S., & Hegel, F. (2012). 'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *2012 7th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 125–126). tex.organization: IEEE.
Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology, 51*(4), 724–731.
Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence, 1*(6), 261–262.
Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines,* 1–14.
Frith, C. D., & Frith, U. (1999). Interacting minds: A biological basis. *Science, 286*(5445), 1692–1695.
Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron, 50*(4), 531–534.
Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: The mirror neuron system responds to human and robotic actions. *Neuroimage, 35*(4), 1674–1684.
Hackel, L. M., Looser, C. E., & Van Bavel, J. J. (2014). Group membership alters the threshold for mind perception: The role of social identity, collective identification, and intergroup threat. *Journal of Experimental Social Psychology, 52,* 15–23.
Haring, K. S. , Silvera-Tawil, D., Matsumoto, Y., Velonaki, M., & Watanabe, K. (2014). Perception of an android robot in Japan and Australia: A cross-cultural comparison. In *International conference on social robotics* (pp. 166–175). (tex.organization: Springer)
Hartzog, W. (2016). Et tu, Android? Regulating dangerous and dishonest robots. *Journal of Human-Robot Interaction, 5*(3), 70–81.
Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*(3), 252–264.

Haslam, N., & Bain, P. (2007). Humanizing the self: Moderators of the attribution of lesser humanness to others. *Personality and Social Psychology Bulletin, 33*(1), 57–68.
Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual Review of Psychology, 65,* 399–423.
Heaven, W. D. (2018). Robot laws. *New Scientist, 239*(3189), 38–41.
Heaven, W. D. (2020). A GPT-3 bot posted comments on Reddit for a week and no one noticed. MIT Technology Review. Retrieved November 24, 2020, from https://www.technologyreview.com/2020/10/08/1009845/a-gpt-3-bot-posted-comments-on-reddit-for-a-week-and-no-one-noticed/
Heider, F. (1983). *The psychology of interpersonal relations*. Psychology Press.
Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology, 57*(2), 243–259.
HLEG. (2021). Ethics guidelines for trustworthy AI - FUTURIUM - european commission. FUTURIUM - European Commission, . Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation
Isaac, A. M., & Bridewell, W. (2017). Why robots need to deceive (and how). *Robot Ethics, 2,* 157–172.
Jacob, P. (2019). Intentionality. In E. N. Zalta (eds.) *The Stanford Encyclopedia of Philosophy (Winter 2019 ed.). Metaphysics Research Lab, Stanford University*. Retrieved from https://plato.stanford.edu/archives/win2019/entries/intentionality/
Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1431), 549–559.
Kahn, P. H., Reichert, A. L., Gary, H. E., Kanda, T., Ishiguro, H., Shen, & S..Gill, B. (2011). The new ontological category hypothesis in human-robot interaction. In*2011 6th*, *ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 159–160). tex.organization: IEEE.
Kaminski, M. E., Rueben, M., Smart, W. D., & Grimm, C. M. (2016). *Averting robot eyes. Md. L. Rev., 76,* 983.
Khalid, S., Deska, J. C., & Hugenberg, K. (2016). eye gaze triggers the ascription of others' minds The eyes are the windows to the mind: Direct eye gaze triggers the ascription of others' minds. *Personality and Social Psychology Bulletin, 42*(12), 1666–1677.
Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition, 26*(2), 169–181.
Kuchenbrandt, D., Eyssel, F., Bobinger, S., & Neufeld, M. (2013). When a robot's group membership matters. *International Journal of Social Robotics, 5*(3), 409–417.
Leviathan, Y., & Matias, Y. (2018). Google duplex: An AI system for accomplishing real-world tasks over the phone. Retrieved from https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation
Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies, 77,* 23–37 (**Publisher: Elsevier**).
Lim, S., & Reeves, B. (2010). Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human-Computer Studies, 68*(1–2), 57–68 (**Publisher: Elsevier**).
Ling, Z., Ma, H., Yang, Y., Qiu, R. C. , Zhu, S. C., & Zhang, Q. (2019). Explaining AlphaGo: Interpreting contextual effects in neural networks. arXiv:1901.02184
Lomas, N. (2018). *Duplex shows Google failing at ethical and creative AI design*. Retrieved May 10, 2018, from https://techcrunch.com/2018/05/10/duplex-shows-google-failing-at-ethical-and-creative-ai-design
Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science, 21*(12), 1854–1862.
Malle, B. F. (2011). Attribution theories: How people make sense of behavior. *Theories in Social Psychology, 23,* 72–95.
Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of experimental social psychology, 33*(2), 101–121.
Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology, 93*(4), 491.
Mandell, A. R. , Smith, M., & Wiese, E. (2017). Mind perception in humanoid agents has negative effects on cognitive processing. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 61, pp. 1585–1589). Number: 1 tex.organization: SAGE Publications.
Marchesi, S., Ghiglino, D., Ciardo, F., Perez-Osorio, J., Baykara, E., & Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology, 10,* 450.

Metz, C. (2016). In two moves, AlphaGo and lee sedol redefined the future. Wired. Retrieved 2016–03-16 from https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future. WIRED.
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267,* 1–38.
Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine, 19*(2), 98–100.
Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103.
Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In: Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 72–78).
Oberman, L. M., Pineda, J. A., & Ramachandran, V. S. (2007). The human mirror neuron system: A link between action observation and social skills. *Social Cognitive and Affective Neuroscience, 2*(1), 62–66.
Ohmoto, Y. , Karasaki, J., & Nishida, T. (2018). Inducing and maintaining the intentional stance by showing interactions between multiple agents. In: *Proceedings of the 18th*, *International Conference on Intelligent Virtual Agents* (pp. 203–210).
O'Leary, D. E. (2019). GOOGLE's duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management, 26*(1), 46–53.
Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C. C., & Feldman, J. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition, 130*(3), 360–379.
Pantelis, P. C., Gerstner, T., Sanik, K., Weinstein, A., Cholewiak, S. A., Kharkwal, G., & Feldman, J. (2016). Agency and rationality: Adopting the intentional stance toward evolved virtual agents. *Decision, 3*(1), 40.
Parkinson, B. (2012). *Social perception and attribution*. Hewstone, M.; Stroebe, W.; Jonas, K. (red.), An Introduction to Social Psychology, 55–90.
Perez-Osorio, J., & Wykowska, A. (2019). Adopting the intentional stance towards humanoid robots. In *Wording robotics* (pp. 119–136). Springer.
Pinchbeck, D. (2008). Trigens can't swim: Intelligence and intentionality in first person game worlds. In: *Proceedings of the*, *Philosophy of Computer Games*, 2008 (pp. 242–260). Potsdam University Press.
Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people.* Cambridge University Press.
Riether, N., Hegel, F., Wrede, B., & Horstmann, G. (2012). Social facilitation with social robots? In: *2012 7th*, *ACM/IEEE International Conference on Human-Robot Interaction* (HRI) (pp. 41–47). tex.organization: IEEE.
Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Reviews Neuroscience, 27,* 169–192.
Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience, 7*(4), 413–422.
Schellen, E., & Wykowska, A. (2019). Intentional mindset toward robots-open questions and methodological challenges. *Frontiers in Robotics and AI, 5,* 139 (**Publisher: Frontiers**).
Searle, J. (1980). Intrinsic intentionality. *Behavioral and Brain Sciences, 3*(3), 450–457.
Seibt, J. (2017). Towards an ontology of simulated social interaction: varieties of the "As If" for robots and humans. In *Sociality and normativity for robots* (pp. 11–39). Springer.
Severson, R. L. & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks, 23*(8–9), 1099–1103.
Shahriari, K,. & Shahriari, M. (2017). IEEE standard review–Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)* (pp. 197–201). tex.organization: IEEE.
Sharkey, N., & Sharkey, A. (2010). The crying shame of robot nannies: An ethical appraisal. *Interaction Studies, 11*(2), 161–190.
Shim, J., & Arkin, R. C. (2012). Biologically-inspired deceptive behavior for a robot. In *International conference on simulation of adaptive behavior* (pp. 401–411). tex.organization: Springer.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., & Guez, A. (2017). Mastering the game of Go without human knowledge. *Nature, 550*(7676), 354–359.

Slors, M. (1996). Why Dennett cannot explain what it is to adopt the intentional stance. *The Philosophical Quarterly, 46*(182), 93–98.
Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds and Machines, 16*(2), 141–161.
Spatola, N., Belletier, C., Chausse, P., Augustinova, M., Normand, A., Barra, V., & Huguet, P. (2019). Improved cognitive control in presence of anthropomorphized robots. *International Journal of Social Robotics, 11*(3), 463–476.
Spatola, N., Monceau, S., & Ferrand, L. (2019). Cognitive impact of social robots: How anthropomorphism boosts performances. *IEEE Robotics & Automation Magazine, 27*(3), 73–83.
Spatola, N., & Normand, A. (2020). Human vs. machine: The psychological and behavioral consequences of being compared to an outperforming artificial agent. *Psychological Research,* 1–11.
Spunt, R. P., Meyer, M. L., & Lieberman, M. D. (2015). The default mode of human brain function primes the intentional stance. *Journal of Cognitive Neuroscience, 27*(6), 1116–1124.
Stich, S. P. (1985). Could man be an irrational animal? Some notes on the epistemology of rationality. *Synthese,* 115–135.
Terada, K., Shamoto, T., Ito, A., & Mei, H. (2007). Reactive, Movements of Non-Humanoid Robots Cause Intention Attribution in Humans. In *2007 IEEE/RSJ international conference on intelligent robots and systems* (pp. 3715–3720). tex.organization: IEEE.
Thellman, S., Silvervarg, A., & Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology,* 8.
Thellman, S., & Ziemke, T. (2019). The intentional stance toward robots: conceptual and methodological considerations. In *The 41st annual conference of the cognitive science society, July 24–26, Montreal, Canada* (pp. 1097–1103).
Theodorou, A., Wortham, R. H., & Bryson, J. J. (2016). Why is my robot behaving like that? Designing transparency for real time inspection of autonomous robots. In: *AISB workshop on principles of robotics*. tex.organization: University of Bath.
Turkle, S. (2010). In good company?: On the threshold of robotic companions. In *Close engagements with artificial companions* (pp. 3–10). Benjamins.
Urgen, B. A., Kutas, M., & Saygin, A. P. (2018). Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia, 114,* 181–185.
Wagner, A. R., & Arkin, R. C. (2011). Acting deceptively: Providing robots with the capacity for deception. *International Journal of Social Robotics, 3*(1), 5–26.
Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219–232.
Weick, K. E. (1995). *Sensemaking in organizations* (Vol. 3). Sage.
Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science, 16*(4), 409–421.
Wiese, E., Buzzell, G. A., Abubshait, A., & Beatty, P. J. (2018). Seeing minds in others: Mind perception modulates low-level social-cognitive performance and relates to ventromedial prefrontal structures. *Cognitive, Affective, & Behavioral Neuroscience, 18*(5), 837–856.
Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology, 8,* 1663.
Willemse, C., Marchesi, S., & Wykowska, A. (2018). Robot faces that follow gaze facilitate attentional engagement and increase their likeability. *Frontiers in Psychology, 9,* 70.
Ziemke, T. (2020). Understanding robots. *Science Robotics, 5,* 46.

CHAPTER 6

Paper5

At the time of the submission of this dissertation, the following paper is still forthcoming in the Proceedings of the "*Robophilosophy 2022: Social Robots in Social Institutions*" conference.

October 2022

Explaining Intentional and Unintentional Behavior: Social Norms for Explainable Robots

Guglielmo PAPAGNI [a,1], Sabine KOESZEGI [a]
[a] *Institute of Management Science, TU Wien, Austria*

Abstract. This paper addresses the question of whether robots should adhere to the same social norms that apply to human-human interaction when they explain their behavior. Specifically, this paper investigates how the topics of ascribing intentions to robots' behavior, and robots' explainability intertwine in the context of social interactions. We argue that robots should be able to contextually guide users towards adopting the most appropriate interpretative framework by providing explanations that refer to intentions, reasons and objectives as well as different kinds causes (e.g., mechanical, accidental, etc.). We support our argument with use cases grounded in real-world applications.

Keywords. Explainable social robots, Intentionality, Explainability

1. Introduction

Recently, interest in explainable artificial intelligence (AI) and robots has been growing, aiming to shed light on AI and robots' otherwise opaque decision-making processes [1-3]. As users' interactions with robots in social contexts will only increase in the future, the understandability, trustworthiness and, ultimately, acceptance of robots at least partially depend on their capacity to explain and justify their behavior [4-6].

In human-human interaction, everyday explanations represent fundamental forms of social communication that people use to transfer knowledge, find meaning regarding the causes of events, influence each other's opinions and manage social interactions [7-10]. People seek explanations primarily when something unexpected, anomalous or abnormal happens [4, 5], which means events or behavior that appear to depart from what social norms would dictate to be the normal and expected course of events in a given situation. People's attribution (or lack thereof) of mental states to other actors, as well as how responsibility and blame are distributed [11, 12], depend on how unexpected behavior and events are explained [7].

In this context, one crucial element concerns whether such events are explained as resulting from 'intentionality', reasons, intentions, beliefs, desires and goals or external and 'unintentional' (natural, mechanical or accidental) factors [4, 12, 13]. Intentional-

[1]Corresponding author: Guglielmo Papagni, Institute of Management Science, TU Wien, Theresianumgasse 27, 1040 Vienna, Austria; e-mail: guglielmo.papagni@tuwien.ac.at

ity is typically interpreted as a byproduct of biological evolution. It represents the relationship between one's internal states (i.e., intentions, desires and beliefs) and the state of things in the world [14, 15]. Despite a certain degree of mutual 'inscrutability' that makes it difficult to infer others' mental states accurately [16, 17], people are accustomed to acknowledging and recognizing each other's intentions (and reasons, goals and beliefs accordingly).

However, with respect to explainable robots, the distinction between 'intentional' and 'unintentional' events is even more blurred and unclear. In fact, even if robots' 'internal architecture' is known, how they compute information represents a very different 'black-box' compared to humans' cognitive processes, which makes it difficult for people to relate to. Perhaps more importantly, being robotic artefacts, they cannot possess 'genuine' or 'intrinsic' intentionality and other mental states. Due to their artificial nature, robots act and make decisions only as a consequence of their programming. They have no actual mental states that relate them to the world and, therefore, no 'genuine intentionality'. Hence, intuitively a mechanistic framework should always suffice to interpret and understand robots' behavior. Accordingly, robots' explanations draw upon design or programming features should be adequate. However, a growing number of studies shows how people tend to attribute intentions and other mental states to robots and other artificial agents regardless of their nature [18, 19]. Furthermore, researchers argue that robots should be aware of and explain their behavior in relation to people's expectations and social norms concerning intentional and unintentional events, or else they may be considered not only erratic and untrustworthy, but also 'immoral' [6, 20, 21].

This paper provides a novel perspective on how to deal with the phenomenon of ascribing intentions to robots' behavior in relation to robots' explainability. After discussing relevant work concerning 'genuine' (or 'intrinsic') intentionality and 'ascribed intentionality' in Section 2, in Section 3 we argue that robots should not only explain events with reference to intentions and other mental states. Rather, because people tend to easily adopt an intentional, mentalistic framework, when needed, robots should also provide explanations that refer to factors and causes of a different nature (e.g., accidental and mechanistic). Furthermore, an observer perceiving a robot's behavior as abnormal and unexpected may imply that the observer is adopting the wrong framework. Hence, we aim to integrate human perception of events into the loop and illustrate, by means of literature-based use cases, how robots should explain events based on these conditions. Section 4 presents final considerations and directions for future work.

2. Related Work

While the idea of attributing intentionality to the behavior of robots and artificial agents has become the object of increasing academic interest recently, it has deep and extensive roots. First, the phenomenon falls within the broader tendency to ascribe anthropomorphic and social attributes to machines [22, 23]. Additionally, study of the topic can be traced back to, at least, work by Heider and Simmel demonstrating how people attribute social significance even to the seemingly 'autonomous' motion of simple geometric shapes (such as triangles, squares and circles) [24].

Much of the recent work on human-robot and human-computer interaction (HRI and HCI respectively) is related to, if not directly inspired by Daniel Dennett's ideas

of 'intentional systems' and the 'intentional stance' [25, 26]. Dennett suggests that it is possible to make sense of certain events by resorting to either the 'physical stance' or the 'design stance' alone. This means that to explain and understand such events, it is enough to appeal to physical laws or an object's function as intended by its design. However, when it comes to sophisticated machines such as robots, due to the very complexity of their inner workings, people can only make sense of their behavior by attributing intentions and rationality to them (i.e., adopting the 'intentional stance') [25-27].

2.1. Instrinsic Intentionality and Ascription of Intentions

The attribution of mental states to robots and computers has subject of extensive debates. One of the most contested points in the formulation of the intentional stance is the underlying 'behaviorism' that Dennett expresses commitment to on several occasions. In fact, while he is aware of the difference between genuinely intentional systems and "those we may find it handy to treat as if they had beliefs and desires" [28: 66], he stresses that manifest behavior is the only way to infer other's mental states, and that behavioral expressions are all that is real. From this perspective, humans are as much 'philosophical zombies' as robots that behave in a seemingly rational manner [29-31]. Then, following up on Putnam's early positions, functionalist and computationalist accounts express the more radical idea of mental states as something independent from biological brains, so that 'zombie robots' may also attain the status of cognitive agents with 'original' intentionality and mental states [32, 33].

In contrast with these views, Searle notes that the attribution of intentionality from an observer's perspective (i.e., 'derived intentionality') should not be confused with 'the real thing' (i.e., 'intrinsic intentionality'). According to Searle, not only do robots and other artefacts not possess intrinsic intentionality, they do not possess any kind of intentionality [34]. Furthermore, referring to Dennett, he observes that "the whole point of his theory of *The Intentional Stance* is to deny that there is any genuine, real, or intrinsic intentionality at all. On Dennett's view there is only *as if* intentionality" [27: 527]. Similarly, Block points out that an artificial entity that looks just like a human could interact just like a real person if adequately pre-programmed and yet lack the very 'essence' (the 'absent qualia') of what it is and feels to be intelligent or have intentions. Lacking the required depth and richness of information processing, seemingly intelligent machines only simulate intelligence and, at best, reveal the intelligence of their programmers [35].

Beyond stressing the fundamental difference between genuine intentions and observers' ascriptions of intentions, Searle also notices how much of the debate seems to be informed by a fundamental misconception regarding this distinction. Interpreting machines' actions as intentional does not necessarily correspond to thinking that machines actually have any intentions [34]. Rather, conflating the two notions may be counterproductive for understanding the phenomenon and its implications [36].

For HRI, this distinction is of fundamental importance for—at least—two reasons. First, as de Graaf and Malle suggest, ascribing intentions and mental states to machines does not necessarily involve attributing any 'self-awareness' or 'consciousness' to them [37]. Rather, the phenomenon can be explained by the social training people are exposed to from an early age, which results in a 'primacy of the social mindset' [38, 39]. The availability of this interpretative framework makes it easy for people to ascribe intentions and mental states but also rationality and subjectivity. This is particularly the case

whenever certain prerequisites (taken alone or combined) are met. These include interaction tasks involving social cognition [38, 40], the level of agency expressed by artificial agents [19], physical embodiment, and anthropomorphism more specifically [36, 41]. Therefore, such an intuitive mechanism may ease social interactions by helping people to make sense of robots' behavior within a familiar human framework [39, 42].

The second reason is that, whatever framework people may find useful to interpret robots' behavior, they fundamentally remain programmed entities, and attributing mind features to them may not always be the best strategy. Particularly, note Weise et al., if a robot's behavior induces 'categorical uncertainties' (i.e., when users cannot tell for sure the nature of the agent), or if it "deviates strongly enough from human behavior so that an anthropomorphic model would lead to incorrect predictions (...) Trying to resolve this cognitive conflict takes up cognitive resources" [41: 8]. Stich identifies the source of the problem in Dennett's understanding of intentional agents, since he posits that intentional agents are agents that always act rationally. However, continues Stich argues, people often behave irrationally. If, according to Dennett's view, intentionality and rationality are necessarily paired, then someone who behaves irrationally cannot be bestowed the status of 'intentional agent' [43].

To summarize, the attribution of intentions and other mental features does not necessarily entail any ontological commitment. Rather, it can be read as a strategy people intuitively adopt to cope with robots' otherwise difficult-to-interpret behavior. Adopting a mentalistic strategy may therefore benefit human-robot interaction but, at least in certain cases, it is important for people to be able to adopt a mechanistic alternative. The next section identifies and discusses four scenarios that require robots to explain events either 'intentionally' or 'unintentionally'. Additionally, as the attribution of intentions to behavior is a subjective experience, the use cases aim to bring human perception into the loop, by considering how subjects may initially perceive and interpret the events in need of explanation.

3. Intentional and Unintentional Explanations

As an alternative to functionalist and behaviorist positions, as well as perspectives that conflate the idea of ascribing intentions with that of 'biological' intentionality, this paper stresses that robots should be able to contextually guide users towards the most appropriate interpretative framework (i.e., mechanistic or mentalistic) for each event that requires an explanation. To illustrate this position, the scenarios discussed further on rest on Heider's idea of 'equifinality' as an antecedent of intentionality [44]. This means that an agent will resort to different means to achieve a goal and that an event may be explained by referring to intentions if, even as the circumstances change, the final objective remains the same. Heider's definition is corroborated by 'folk-psychological' readings of intentionality as a social construct, a tool that people use not only to recognize (biological) conspecifics, but also to correctly interpret and predict how others (both humans and not) act [36], as well as to facilitate social interactions [45]. Hereafter, when we describe robots' behavior as 'intentional', this means that a specific action is aligned with a robot's overall purpose, rather than implying any actual mental state behind it.

Furthermore, the argument that an observer can only infer others' (humans or robots) intentions should in principle also be applicable to behavior that is best interpreted as

unintentional (e.g., if a robot stops carrying out its task because of an obstruction). To cope with this (double) limitation in the actor-observer interaction, robots "must be able to distinguish intentional from unintentional behaviors" and, at the same time, they "must be able to explain each of these classes of behavior in the expected way—unintentional behaviors with (mere) causes, intentional behaviors with reasons" [37: 19].

We identify two main reasons that support this argument. First, while an intentional or 'mentalistic' interpretation may actually be preferable in many interaction instances, such as tasks that involve social cognition and, specifically, when a robot's behavior displays 'equifinality', not all events can and should be explained with respect to intentions and other mental states. To the contrary, designing robots to explain certain events and behavior with respect to accidental, natural or mechanistic causes is particularly important precisely because, as previously shown, people tend to easily adopt an intentional framework, even when it's not the most appropriate one.

The second reason refers to the fact that explanations are mostly requested when something unpredictable or unexpected happens. A robot's behavior may be perceived as unexpected simply because users cannot immediately and correctly ascribe reasons and intentions in relation to what they perceive the robot's goal to be. Alternatively, the possibility exists that the interpretative framework a user has adopted is not the most appropriate one to interpret the event that triggers the explanation request. In this regard, if users cannot make sense of the reasons, intentions or causes behind robots' behavior, users' understanding and trust will likely be at stake [4-6]. If adequately tailored, explanations can not only avoid or mitigate trust losses, they may also prevent users predisposed to having high expectations from over-trusting robots [4, 46]. Explanations' success in calibrating users' expectations and trust inevitably depends, among other things, on robots' capability to guide users towards the most appropriate framework for each interaction context.

3.1. Four Use Cases for Explainable Robots

Thellman and Ziemke suggest that only events such as a robot running out of battery should be explained by adopting a more mechanistic approach (or the 'design stance') [47]. However, this is a reduction to an ideal and simplified scenario, while everyday interactions typically pose more nuanced challenges. Furthermore, we argue that a dichotomy, with events labeled only as either 'intentional' or 'unintentional', is not sufficient. Whether a robot's behavior is considered the result of intentions or of other causes does not only depend on the robot's actual action plan, but also on how observers perceive the event in regards to their expectations and contingent social norms. In other words, a specific action by a robot that is actually aligned with its plan may be initially perceived by users as accidental, or vice versa. In this regard, Bartneck and Forlizzi argue that social robots should be aware of the knowledge about the world they possess as well as, whenever possible, of what they do not know [21]. Hence, the value of our argument is found particularly in the grey areas, as these signify that users may need to switch from one interpretation to another.

It is important to note that the need for guidance regarding the most appropriate framework is likely to be particularly fundamental (but not exclusively so) in interactions with robots 'in the wild'. In fact, these kinds of encounters are likely to be 'one-time' events, mostly involving laypeople who have little to no chance of undergoing specific

training and preparation for the robots they encounter. This lack of previous experience implies that people will often not know what to expect from robots, making it more difficult to infer what should be considered 'intentional' and what should not.

The next paragraphs present four interaction scenarios that require explanations, in which a robot's behavior is:

- Intentional and (correctly) interpreted through an intentional framework;
- Unintentional and (erroneously) interpreted through an intentional framework;
- Intentional and (erroneously) interpreted through an unintentional framework;
- Unintentional and (correctly) interpreted through an unintentional framework;

3.1.1. Shopping for Books with a Robot

We previously noted how unexpected events are the main trigger for explanation requests. However, people may have other reasons, such as curiosity about how a robot generates specific decisions or suggestions, to ask for an explanation [1]. Social robots are employed as shopping assistants in book stores, shopping malls and other social spaces [48, 49]. Here, a customer may want to know, for instance, the reasons for a specific book recommendation because they are not familiar with the author or the title, and therefore ask for an explanation. Importantly, as long as the customer is aware of the robot's role in the store (i.e., to assist people with their purchases), the robot's behavior of recommending specific books in accordance with customers' preferences would likely not be considered unusual or a violation of social norms. Rather, the customer would likely treat the robot's recommendation as intentional behavior and yet still be interested to know how the robot came up with that specific recommendation (perhaps before deciding whether to buy the book or not). Accordingly, the robot may explain that the recommendation was (intentionally) generated on the basis of the customer's previous purchases and books' ratings, combined with those of other customers with similar tastes. In this specific case, the intentional framework emerges as the most adequate to make sense of the robot's behavior and no switch is required or should be promoted by the robot's explanation.

Unintentional Recommendation The previous case is also helpful to understand the second scenario, in which unintentional behavior is initially interpreted as intentional. Importantly, Heider's original argument on 'equifinality' involves repeated observations to establish whether or not to attribute intentionality to a specific act or behavior [44]. Considering the same customer going back to the book store to look for more books, one can assume that, after the first interaction, they would have likely formed an idea of how the robot operates to reach its intended purpose. It is important to stress that the robot's intended purpose is not to provide 'just any' book recommendations, as this would be equal to picking out books randomly. Rather, it is to provide 'individually tailored', useful recommendations. As in the previous scenario,ifs the robot suggests a book unknown to the customer, they may want to know how the robot came up with that specific recommendation, initially assuming that the robot's behavior is 'normal' and aligned with its plan (i.e., intentional), as in the previous interaction. The robot would then show how various features were weighed in generating the book recommendation, with the difference that, in this case, the features do not match the customer's actual data, but rather those of another person. At this point, the customer's expectations have likely been violated, as they realize that the recommendation is not accurate. To help the

customer avoid a 'cognitive conflict', when they ask for further clarification, the robot should support them in switching to a mechanistic framework to interpret the 'wrong' recommendation. For instance, the robot could explain that a mistake in data processing may have occurred and that the last recommendation was not aligned with its intended purpose.

3.1.2. Medicines Delivery Robots

One area of application for assistive robots is in elderly care contexts (at home, or in nursing facilities) with tasks that include monitoring people's health, reminding them of and delivering scheduled medicines [50]. In a similar context, the task of reminding patients and residents to take medicine is meant to support people affected by old age-related memory loss. For instance, a robot may have to remind a user to take their medicine every day after meals. However, if the user has forgotten this schedule, they may find the behavior abnormal, wonder whether the robot made a scheduling mistake, and ask for an explanation accordingly. At this point, the robot should help the user remember by clarifying that they are supposed to take their medicines after every meal. In other words, by clarifying that it was acting in accordance with plan of reminding and delivering medicines according to a prescribed schedule, the robot should guide the user's interpretation from an unintentional framework (i.e., the robot made a mistake) to an intentional one.

3.1.3. Pick up Failure

Finally, there are situations in which users will likely see immediately that a robot's behavior is caused by external factors and should not be interpreted as intentional. Referring to the previous case of a service robot reminding patients of and delivering medicines, it may happen that, as the robot approaches a user, it fails to pick up and hand over the medicines. The user would likely realize quickly that the robot's behavior (failing to pick up) is not intentional, as it is not in any way useful to reach the robot's intended purpose (delivering the medicine). Yet, they might ask the robot to explain why it could not pick up the medicine, to figure out whether to call for help or not. While it might not be possible for the robot to identify the causes with certainty (or else it would have likely succeeded), it could nevertheless explain with a certain degree of confidence that an object may be obstructing its view of the medicines to deliver, making it difficult to correctly estimate their position and, consequently, pick them up. In other words, the robot should emphasize the accidental causes and let users know that no switch is required, as a mechanistic framework is the most appropriate to make sense of the event.

As a closing remark, it is important to recall that people and robots can only infer others' intentions or lack thereof. Concerning explanations, we argued that robots should help users adopt the best interpretative framework. However, in light of both the general inaccessibility of others' minds, and robots currently limited capabilities, they may not always be able to correctly infer whether people adopt the most appropriate framework at specific points of an interaction. Hilton notes that a good social explanation is one that provides context-relevant information [1]. Additionally, when providing explanations, people try to adapt them to the explainee in terms of complexity, amount of information, and clarity [51]. Following these principles, and given the nuances of everyday interactions, what robots can do is try to guide users towards adopting the most appropriate

framework in each situation, particularly when it may be unclear whether an event or behavior should be interpreted as intentional or not. This can be done by tailoring explanations to contextual conditions and implementing specific strategies that improve their quality. In this regard, studies show that explanations' multi-modality, interactivity and questionability represent promising strategies that have not yet been fully investigated [52, 53]. For instance, when reminding a user to take their medicines, a robot may corroborate a text-based explanation with a graphic visualization of the schedule prescribed by a doctor.

4. Conclusions

This paper discussed how social robots should provide explanations in relation to attribution of intentions. We argued that, while robots have no 'genuine intentionality' per se, in many instances, it may be helpful for users to interpret robots' behavior 'as if' it resulted from intentions and other mental states. Ascribing intentions is an easily triggered cognitive mechanism which may help users interpret robots' otherwise inscrutable inner workings.

However, precisely because it is so easy for people to ascribe intentions to others' behavior (whether human or not), certain situations that may be better interpreted as caused by external, non-intentional factors will require robots to support users in adopting a mechanistic framework by means of explanations. At the same time, the opposite scenario (i.e., that robots may have to guide users towards an intentional interpretation) could also occur. Through use cases based on real-world applications of social robots, we illustrated how, in general, the intrinsic nuances of everyday interactions require robots' explanations to be finely tuned to guiding users towards adopting the most appropriate interpretative framework for each situation.

The nature of the arguments this paper proposes are fundamentally conceptual, albeit discussed with the support of use cases inspired by real-world applications. Therefore, the main direction for future work is to test how the nuances of robots' behavior are interpreted by users in terms of intentions or lack thereof. At the same time, how robots' explanations may support users' adoption of the most appropriate framework and how this kind of explanations affect users' trust in the robots must also be empirically assessed.

References

[1] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access. 2018 Sep 17;6:52138-60.

[2] Anjomshoae S, Najjar A, Calvaresi D, Främling K. Explainable agents and robots: Results from a systematic literature review. In Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019); 2019 May 13–17; Montreal, Canada: International Foundation for Autonomous Agents and Multiagent Systems c2019. p. 1078-1088.

[3] Wallkötter S, Tulli S, Castellano G, Paiva A, Chetouani M. Explainable embodied agents through social cues: a review. ACM Transactions on Human-Robot Interaction (THRI). 2021 Jul 10;10(3):1-24.

[4] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence. 2019 Feb 1;267:1-38.

[5] Andras P, Esterle L, Guckert M, Han TA, Lewis PR, Milanovic K, Payne T, Perret C, Pitt J, Powers ST, Urquhart N. Trusting intelligent machines: Deepening trust within socio-technical systems. IEEE Technology and Society Magazine. 2018 Dec 4;37(4):76-83.
[6] Lomas M, Chevalier R, Cross EV, Garrett RC, Hoare J, Kopack M. Explaining robot actions. In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction 2012 Mar 5 p. 187-188.
[7] Hilton DJ. Conversational processes and causal explanation. Psychological Bulletin. 1990 Jan;107(1):65.
[8] Lewis D. Causal Explanation. Philosophical Papers. 1986:214-40.
[9] Malle BF. Attribution theories: How people make sense of behavior. Theories in social psychology. 2011;23:72-95.
[10] Malle BF. How the mind explains behavior: Folk explanations, meaning, and social interaction. MIT press; 2006 Aug 11.
[11] Malle BF, Knobe J. Which behaviors do people explain? A basic actor–observer asymmetry. Journal of Personality and Social Psychology. 1997 Feb;72(2):288.
[12] Knobe J. The concept of intentional action: A case study in the uses of folk psychology. Philosophical studies. 2006 Aug;130(2):203-31.
[13] Malle BF. Folk explanations of intentional action. Intentions and intentionality: Foundations of social cognition. 2001:265-86.
[14] Searle J. Intrinsic intentionality. Behavioral and Brain Sciences. 1980 Sep;3(3):450-7.
[15] Jacob P. Intentionality. Stanford Encyclopedia of Philosophy. 2008.
[16] Avramides A. Other Minds. In Brian McLaughlin, Ansgar Beckermann & Sven Walter (eds.), The Oxford Handbook of Philosophy of Mind. Oxford University Press. 2009.
[17] Malle BF, Knobe JM, Nelson SE. Actor-observer asymmetries in explanations of behavior: new answers to an old question. Journal of personality and social psychology. 2007 Oct;93(4):491.
[18] De Graaf, MM, & Malle, BF. People's explanations of robot behavior subtly reveal mental state inferences. In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) 2019 Mar p. 239-248.
[19] Pantelis PC, Gerstner T, Sanik K, Weinstein A, Cholewiak SA, Kharkwal G, Wu CC, Feldman J. Agency and rationality: Adopting the intentional stance toward evolved virtual agents. Decision. 2016 Jan;3(1):40.
[20] de Graaf MM, Malle BF, Dragan A, Ziemke T. Explainable robotic systems. Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction; 2018 Mar 1. p. 387-388.
[21] Bartneck C, Forlizzi J. A design-centred framework for social human-robot interaction. In13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2004) 2004 p. 591-594.
[22] Nass C, Steuer J, Tauber ER. Computers are social actors. In Proceedings of the SIGCHI conference on Human factors in computing systems 1994 Apr 24 p. 72-78.
[23] Caporael LR. Anthropomorphism and mechanomorphism: Two faces of the human machine. Computers in human behavior. 1986 Jan 1;2(3):215-34.
[24] Heider F, Simmel M. An experimental study of apparent behavior. The American journal of psychology. 1944 Apr 1;57(2):243-59.
[25] Dennett DC. Intentional systems. The Journal of Philosophy. 1971 Jan 1;68(4):87-106.
[26] Dennett DC. The intentional stance. MIT press; 1987.
[27] Dennett DC. Précis of the intentional stance. Behavioral and brain sciences. 1988 Sep;11(3):495-505.
[28] Dennett DC. True Believers: The Intentional Strategy and Why It works. Mind Design. 1997:57-79.
[29] Dennett DC. The unimagined preposterousness of zombies. Journal of Consciousness Studies. 1995;2(4).
[30] Dennett DC. Real Patterns. The Journal of Philosophy. 1991 Jan;88(1):27-51.
[31] Dennett DC. Consciousness Explained. Philosophy and Phenomenological Research. 1993;53(4).
[32] Horgan T. Original intentionality is phenomenal intentionality. The Monist. 2013 Apr 4;96(2):232-51.
[33] Rescorla M. Computational modeling of the mind: what role for mental representation?. Wiley Interdisciplinary Reviews: Cognitive Science. 2015 Jan;6(1):65-73.
[34] Searle JR. Minds, brains, and programs. Behavioral and brain sciences. 1980 Sep;3(3):417-24.
[35] Block N. Psychologism and behaviorism. The Philosophical Review. 1981 Jan 1;90(1):5-43.
[36] Thellman S, Silvervarg A, Ziemke T. Folk-psychological interpretation of human vs. humanoid robot

behavior: Exploring the intentional stance toward robots. Frontiers in psychology. 2017 Nov 14;8:1962.
[37] De Graaf MM, Malle BF. How people explain action (and autonomous intelligent systems should too). In 2017 AAAI Fall Symposium Series 2017 Oct 9.
[38] Perez-Osorio J, Wykowska A. Adopting the intentional stance towards humanoid robots. In Wording robotics 2019 (pp. 119-136). Springer, Cham.
[39] Papagni G, Koeszegi S. A Pragmatic Approach to the Intentional Stance Semantic, Empirical and Ethical Considerations for the Design of Artificial Agents. Minds and Machines. 2021 Dec;31(4):505-34.
[40] Spunt RP, Meyer ML, Lieberman MD. The default mode of human brain function primes the intentional stance. Journal of cognitive neuroscience. 2015 Jun 1;27(6):1116-24.
[41] Wiese E, Metta G, Wykowska A. Robots as intentional agents: using neuroscientific methods to make robots appear more social. Frontiers in psychology. 2017 Oct 4;8:1663.
[42] Breazeal C, Scassellati B. How to build robots that make friends and influence people. In Proceedings of the 1999 IEEE/RSJ international conference on intelligent robots and systems. Human and environment friendly robots with high intelligence and emotional quotients (cat. No. 99CH36289) 1999 Oct 17 Vol. 2, p. 858-863.
[43] Stich SP. Could man be an irrational animal? Some notes on the epistemology of rationality. Synthese. 1985 Jul 1:115-35.
[44] Heider F. The psychology of interpersonal relations. Psychology Press; 2013 May 13.
[45] Malle BF, Knobe J. The folk concept of intentionality. Journal of experimental social psychology. 1997 Mar 1;33(2):101-21.
[46] Lockey S, Gillespie N, Holm D, Someh IA. A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. In Proceedings of the 54th Hawaii International Conference on System Sciences 2021 Jan 5 p. 5463-5472.
[47] Thellman S, Ziemke T. The Perceptual Belief Problem: Why Explainability Is a Tough Challenge in Social Robotics. ACM Transactions on Human-Robot Interaction (THRI). 2021 Jul 10;10(3):1-5.
[48] Sreejith MS, Joy S, Pal A, Ryuh BS, Kumar VS. Conceptual design of a wi-fi and GPS based robotic library using an intelligent system. International Journal of Computer and Information Engineering. 2015 Nov 5;9(12):2504-8.
[49] Niemelä M, Heikkilä P, Lammi H. A social service robot in a shopping mall: expectations of the management, retailers and consumers. In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI) 2017 Mar 6 p. 227-228.
[50] Pollack ME, Brown L, Colbry D, Orosz C, Peintner B, Ramakrishnan S, Engberg S, Matthews JT, Dunbar-Jacob J, McCarthy CE, Thrun S. Pearl: A mobile robotic assistant for the elderly. InAAAI workshop on automation as eldercare 2002 Aug (Vol. 2002).
[51] Cawsey A. User modelling in interactive explanations. User Modeling and User-Adapted Interaction. 1993 Sep;3(3):221-47.
[52] Alipour K, Schulze JP, Yao Y, Ziskind A, Burachas G. A study on multimodal and interactive explanations for visual question answering. arXiv preprint arXiv:2003.00431. 2020 Mar 1.
[53] Park DH, Hendricks LA, Akata Z, Rohrbach A, Schiele B, Darrell T, Rohrbach M. Multimodal explanations: Justifying decisions and pointing to the evidence. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 p. 8779-8788.

Bibliography

[1] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[2] 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. Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems. *IEEE Technology and Society Magazine*, 37(4):76–83, 2018.

[3] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. *Explainable Agents and Robots: Results from a Systematic Literature Review*. International Foundation for Autonomous Agents and Multiagent Systems, May 2019.

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

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

[6] RL Buckner, JR Andrews-Hanna, and DL Schacter. The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124:1–38, 2008.

[7] Alison Cawsey. User modelling in interactive explanations. *User Modeling and User-Adapted Interaction*, 3(3):221–247, 1993.

[8] Shih-Yi Chien, Katia Sycara, Jyi-Shane Liu, and Asiye Kumru. 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*, volume 60, pages 841–845. SAGE Publications Sage CA: Los Angeles, CA, 2016.

[9] John Danaher. Robot betrayal: a guide to the ethics of robotic deception. *Ethics and Information Technology*, pages 1–12, 2020.

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

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

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

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

[14] Daniel C Dennett. *Brainstorms: Philosophical essays on mind and psychology*. MIT press, 2017.

[15] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

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

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

[18] Randi A Engle. 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*, pages 321–326, 1998.

[19] Fabio Fossa. " i don't trust you, you faker!" on trust, reliance, and artificial agency. 2019.

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

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

[22] Denis J Hilton, John McClure, and Robbie M Sutton. Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, 40(3):383–400, 2010.

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

[24] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635, 2021.

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

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

[27] Frank C Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254, 2006.

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

[29] Himabindu Lakkaraju and Osbert Bastani. " 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*, pages 79–85, 2020.

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

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

[32] D LEWIS. Causal explanation. *Philosophical Papers*, pages 214–240, 1986.

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

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

[35] Niklas Luhmann. *Trust and power*. John Wiley & Sons, 2018.

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

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

[38] Bertram F Malle. Attribution theories: How people make sense of behavior. *Theories in social psychology*, 23:72–95, 2011.

[39] Bertram F Malle and Joshua Knobe. The folk concept of intentionality. *Journal of experimental social psychology*, 33(2):101–121, 1997.

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

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

[42] Peter C Pantelis, Timothy Gerstner, Kevin Sanik, Ari Weinstein, Steven A Cholewiak, Gaurav Kharkwal, Chia-Chien Wu, and Jacob Feldman. Agency and rationality: Adopting the intentional stance toward evolved virtual agents. *Decision*, 3(1):40, 2016.

[43] Guglielmo Papagni and Sabine Koeszegi. A pragmatic approach to the intentional stance semantic, empirical and ethical considerations for the design of artificial agents. *Minds and Machines*, 31(4):505–534, 2021.

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

[45] Jairo Perez-Osorio and Agnieszka Wykowska. Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, 33(3):369–395, 2020.

[46] Julian B Rotter. Generalized expectancies for interpersonal trust. *American psychologist*, 26(5):443, 1971.

[47] Cynthia Rudin. Please Stop Explaining Black Box Models for High Stakes Decisions. *arXiv*, Nov 2018.

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

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

[50] F David Schoorman, Roger C Mayer, and James H Davis. An integrative model of organizational trust: Past, present, and future. *Academy of Management review*, 32(2):344–354, 2007.

[51] John Searle. Intrinsic intentionality. *Behavioral and Brain Sciences*, 3(3):450–457, 1980.

[52] John R Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3):417–424, 1980.

[53] Raymond Sheh. Different xai for different hri. In *AAAI Fall Symposium-Technical Report*, pages 114–117, 2017.

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

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

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

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

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

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

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

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

[62] Daniel A Wilkenfeld and Tania Lombrozo. Inference to the best explanation (ibe) versus explaining for the best inference (ebi). *Science & Education*, 24(9-10):1059–1077, 2015.

[63] Kewen Wu, Yuxiang Zhao, Qinghua Zhu, Xiaojie Tan, and Hua Zheng. A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type. *International Journal of Information Management*, 31(6):572–581, 2011.