AI & SOCIETY
https://doi.org/10.1007/s00146-022-01462-7

NETWORK

Artificial agents' explainability to support trust: considerations on timing and context

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

Received: 26 May 2021 / Accepted: 13 April 2022

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 agent'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

1 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

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

Jesse de Pagter
jesse.de.pagter@tuwien.ac.at

Setareh Zafari
setareh.zafari@tuwien.ac.at

Michael Filzmoser
michael.filzmoser@tuwien.ac.at

Sabine T. Koeszegi
sabine.koeszegi@tuwien.ac.at

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

Published online: 27 June 2022

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, Sect. 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 Sect. 3, the main propositions developed throughout the paper are graphically rendered in the form of a model that describes the connections between explanations and trust. Section 4 concludes the study and discusses directions for future research.

2 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.

2.1 Fundamental features of trust

2.1.1 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.

2.1.2 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 agent 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.

2.1.3 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 agent 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).

2.2 (Initial) trust establishment

A potential issue emerges here that concerns initial trust. In fact, on 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.

2.2.1 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.

2.2.2 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 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.

3 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.

3.1 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 other's 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.

3.1.1 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).

Therefore, 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.

3.2 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.

3.2.1 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 where 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 was 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.

3.2.2 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 the agent behaved 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 (see Fig. 1).

4 Communicating explanations

We previously noted how explanations come with at least two major limitations. On 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 'oneshot' 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).

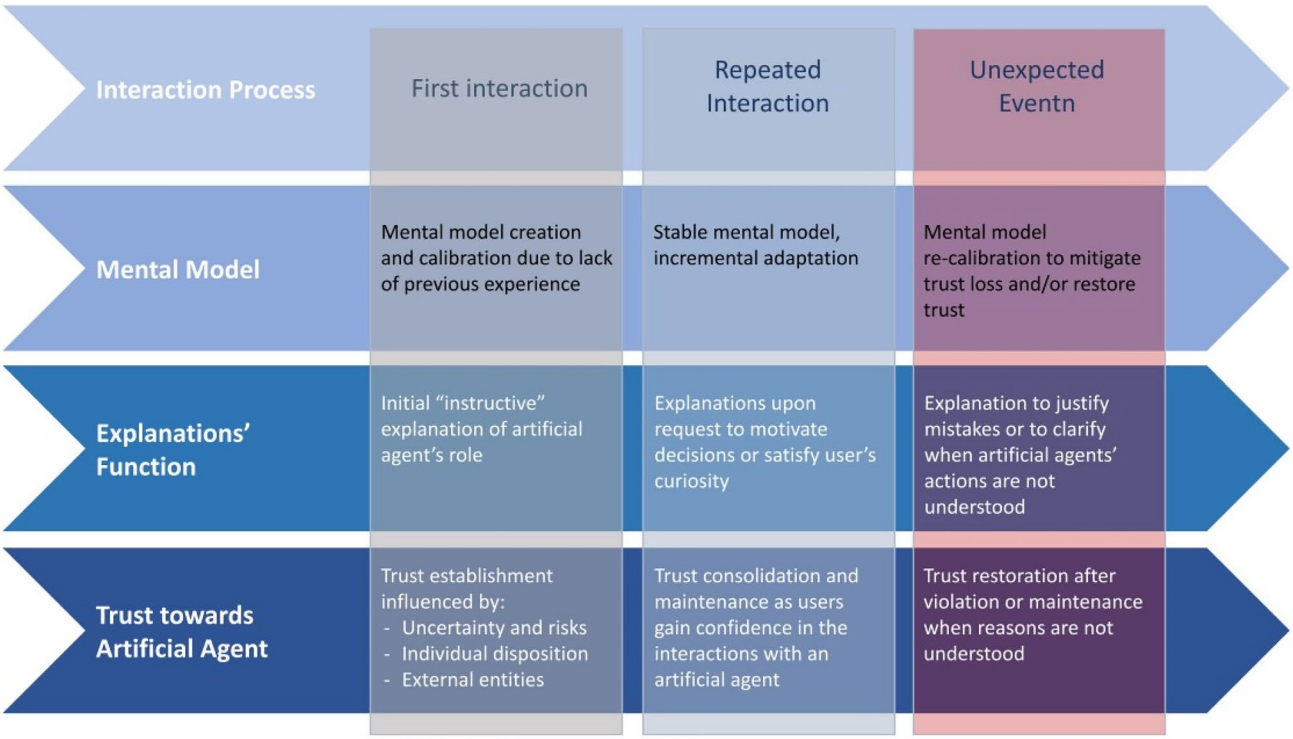

Fig. 1 Graphic visualization of explanations as trust support strategy, throughout repeated interactions

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.

4.1 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 did not 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 gives 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 explanation 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.

4.2 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.

4.3 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 combinamtion of interactive and multimodal 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.

5 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.

Acknowledgements Funding for this study was contributed by MercedesBenz AG. The authors would like to thank Alischa Rosenstein, Dimitra TheofanouFu¨lbier and Joana Hois for their guidance and feedback throughout this project.

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

References

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/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 Technol Soc Mag 37(4):76–83. https://doi.org/10.1109/MTS.2018.2876107

Anjomshoae S, Najjar A, Calvaresi D, Främling K (2019) Explainable Agents and Robots: Results from a Systematic Literature Review. In: Proceedings of the 18th InternationalConference on Autonomous Agents and MultiAgent Systems, pp 1078–1088

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. Sci Educ 93(1):26–55

Cawsey A (1993) User modelling in interactive explanations. User Model User-Adap Inter 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 Inf Technol 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, pp 19–26

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

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. Int J Soc Robot 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. Int J Hum Comput Stud 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, 22(1), 4–24

Elia J (2009) Transparency rights, technology, and trust. Ethics Inf Technol 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.Teoria, 1:63–80

Fulmer CA, Gelfand MJ (2012) At what level (and in whom) we trust: trust across multiple organizational levels. J Manag 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 Mag 38(3):50–57

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 Mag 40(2):44–58

Hagras H (2018) Toward human-understandable. Explainable AI. Computer 51(9):28–36. https://doi.org/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. Hum Factors 53(5):517–527

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

Hellström T, Bensch S (2018) Understandable robots - What, Why, and How. Paladyn J Behav Robot 9(1):110–123. https://doi.org/10.1515/pjbr2018-0009

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. Psychol Bull 107(1):65

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

Hilton DJ, McClure J, Sutton RM (2010) Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? Eur J Soc Psychol 40(3):383–400

Hleg A (2019) Ethics guidelines for trustworthy AI, Publications Office. https://data.europa.eu/doi/10.2759/346720

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-gas acceptance and trust: First-year results and implications. J Cogn Eng Decis 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 https://doi.org/10.1007/978-3-030-23528-425

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. Inf Manag 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 Cogn Sci 7(8):368–373
Kerschner C, Ehlers MH (2016) A framework of attitudes towards technology in theory and practice. Ecol Econ 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. J Assoc Inf Syst 16(10):1
Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Factors 46(1):50–80
Leimstoll U, Stormer H (2007) Collaborative recommender systems for online shops. In: 13th Americas Conference on Information Systems, AMCIS 2007, Keystone, Colorado, USA, August 9–12, 2007
Lewicki RJ, Brinsfield C (2017) Trust repair. Annu Rev Organ Psych Organ Behav 4:287–313
Lewis D (1986) Causal Explanation. Philosophical Papers Vol ii, Oxford University Press, 214-240
Li X, Hess TJ, Valacich JS (2008) Why do we trust new technology? A study of initial trust formation with organizational information systems. J Strateg Inf Syst 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 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, pp 5463–5472
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 Cogn Sci 10(10):464–470. https://doi.org/10.1016/j.tics.2006.08.004
Lombrozo T (2007) Simplicity and probability in causal explanation. Cogn Psychol 55(3):232–257
Luhmann N (2000) Familiarity, confidence, trust: problems and alternatives. Trust Making Break Cooper Relat 6(1):94–107
Luhmann N (2018) Trust and power. Wiley
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. Theor Socl Psychol 23:72–95
Malle BF, Knobe JM, Nelson SE (2007) Actor-observer asymmetries in explanations of behavior: New answers to an old question. J Pers Soc Psychol 93(4):491
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. Artif Intell 267:1–38. https://doi.org/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. Pers Psychol 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. Intell Syst Account Finance Manag 26(1):46–53
O'neill O (2002) Autonomy and trust in bioethics. Cambridge University Press, Berlin
Papagni G, Koeszegi S (2020) Understandable and trustworthy explainable robots: A sensemaking perspective. Paladyn J Behav Robot 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. Mind Mach 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 Inf Technol 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. Adv Exp Soc Psychol 41:1–67
Pu P, Chen L (2007) Trust-inspiring explanation interfaces for recommender systems. Knowl-Based Syst 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 Behav Emerg Technolos 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 Trans HumanMach Syst 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. Am Psychol 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 141–148

Schoorman FD, Mayer RC, Davis JH (2007) An integrative model of organizational trust: past, present, and future. Acad Manag Rev 32(2):344–354

Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. Int J Human-Comput Stud 146:102551

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

Simpson JA (2007) Foundations of interpersonal trust. Soc Psychol Handb Basic Princ 2:587–607

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

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

Thagard P (1989) Explanatory coherence. Behav Brain Sci 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

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

Venkatesh V, Morris MG, Ackerman PL (2000) A longitudinal field investigation of gender differences in individual technology adoption decision-making processes. Organ Behav Hum Decis Process 83(1):33–60

Wagner AR, Borenstein J, Howard A (2018) Overtrust in the robotic age. Commun 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. J Mach Learn Res 22(137):1–38

Weick KE, Sutcliffe KM, Obstfeld D (2005) Organizing and the process of sensemaking. Organ Sci 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). Sci Educ 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. J Med Internet Res 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. Organ Sci 9(2):141–159

Zemla JC, Sloman S, Bechlivanidis C, Lagnado DA (2017) Evaluating everyday explanations. Psychon Bull Rev 24(5):1488–1500

Zou J, Schiebinger L (2018) AI can be sexist and racist—it's time to make it fair, 559(7714):324–326

Zucker LG (1987) Institutional theories of organization. Ann Rev Sociol 13(1):443–464