The Effect of Explanation Styles on User's Trust

Retno Larasati
retno.larasati@open.ac.uk
Knowledge Media Institute
The Open University
UK

Anna De Liddo
anna.deliddo@open.ac.uk
Knowledge Media Institute
The Open University
UK

Enrico Motta
enrico.motta@open.ac.uk
Knowledge Media Institute
The Open University
UK

ABSTRACT

This paper investigates the effects that different styles of textual explanation have on explainee's trust in an AI medical support scenario. From the literature, we focused on four different styles of explanation: contrastive, general, truthful, and thorough. We conducted a user study in which we presented explanations of a fictional mammography diagnosis application system to 48 non-expert users. We carried out a between-subject comparison between four groups of 11-13 people each looking at a different explanation style. Our findings suggest that contrastive and thorough explanations produce higher personal attachment trust scores compared to general explanation style, while truthful explanation shows no difference compared to the rest of explanations. This means that users who received contrastive and thorough explanation types found the explanation given significantly more agreeable and suiting their personal taste. These findings, even though not conclusive, confirm the impact of explanation style on users trust towards AI systems and may inform future explanation design and evaluation studies.

CCS CONCEPTS

• **Human-centered computing** → *Human Computer Interaction.*

KEYWORDS

Explanation, Trust, Explainable Artificial Intelligence

ACM Reference Format:
Retno Larasati, Anna De Liddo, and Enrico Motta. 2020. The Effect of Explanation Styles on User's Trust. In *Proceedings of IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies (ExSS-ATEC'20).* , 6 pages.

1 INTRODUCTION

One of the main arguments motivating Explainable Artificial Intelligence research is that the explicability of AI systems can improve people's trust and adoption of AI solutions [8][27]. Still, the relationships between trust and explanation is complex, and it is not always the case that explicability improves users' trust. Trust in AI systems is claimed to be enhanced by transparency [11] and understandability [17]. In order to gain understandability, an AI system should provide explanations that are meaningful to the explainee(someone who received explanation). Providing meaningful explanations could then support users to appropriately calibrate trust, by improving trust (when they tend to down-trust the system) and mitigating over-trust issues [31]. Previous research has shown that a key role in calibrating trust can be played by the way in which explanation is expressed and presented to the users. Explanation style and modalities affect users' trust toward algorithmic systems sometime improving sometime reducing trust [12][25]. This paper aims to investigate the relation between explanation and trust by exploring different explanation styles. We first conducted a literature review in psychology, philosophy, and information systems, to understand what are the characteristics of meaningful explanations. We then designed several styles of explanation based on these characteristics. Since we are interested in assessing the effects of explanation styles on users' trust, we also defined a variety of trust components to measure users' trust levels. Our proposed trust measurement was gathered from the literature in human factors and HCI research. Finally we carried out a user study to see if any specific explanation style differently affects users' trust. Our contribution is twofold:

(1) we provide evidence which confirms the effect of explanation styles on different trust factors;
(2) we propose a reliable human-AI trust measurement (Cronbach's α=0.88) to investigate explanation and trust in healthcare.

Thie rest of the paper is organized as follows: Section 2 introduces the context of this research and summarises the relevant literature. Section 3 describes the methodology of this study. Section 4 and 5 presents and analyse the results from the study. Finally, Section 6 discusses the limitation of this work and outlines the next steps of the research.

2 BACKGROUND AND RELATED WORK

2.1 Explanation

Explanation can be seen as an act or a product and can be categorised as good or bad. A good explanation is an explanation that feels right because offers a phenomenologically familiar sense of understanding [1]. In this paper, we focus on meaningful explanation, to stress our interest and focus on the explanation's capability to improve understanding and sense-making of AI and algorithmic results. As such good explanations are not explanations that necessarily improve trust, and can affect user's trust both ways, by either improving or moderating trust.

We might ask what is meaningful explanation? There is no single definition of meaningful explanation. Guidotti et al. defined meaningful explanation as explanation that is faithful and interpretable [7]. Thirumuruganathan et al. defined meaningful explanation as explanation that is personalised based on users' demographic [29]. Regulators have also mentioned meaningful explanation. GDPR Articles 13–15 state that users have the right to receive 'meaningful information about the logic involved' in automated decisions, but

ExSS-ATEC'20, March 2020, Cagliari, Italy,

it fails to provide any specific definition of what is to be considered 'meaningful information'. In this paper, we will refer to meaningful explanation as explanation that is understandable.

In cognitive psychology, explanation can be classified into different types: i. Causal explanation, which tells you what causes what, ii. Mechanical explanation, which tells you how a certain phenomenon comes about, and iii. Personal explanation which tells you what causes what in the context of personal reasons or beliefs [32]. Approaching these definitions from an explainable AI and AI reasoning angle, we could say that causal and mechanical explanation could be the same, because the causal explanation of an AI system is mechanical by definition. For instance, if we ask why the AI system gives us a certain prediction, the answer will consist of an illustration of the AI's mechanical process, which produced that prediction result. Personal explanation might also not be relevant, since all AI "personal" explanations are defined in terms of what causes what in the context of a specific AI reasoning mechanism. Therefore, in what follows we will focus on causal explanation.

Hilton proposed that causal explanation proceeds through the operation of counterfactual and contrastive criteria [10]. Lipton suggested that *"to explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q"* [16]. Miller quoted Lipton and argued that everyday explanations, or human explanations, are *"sought in response to particular counterfactual cases. [...]people do not ask why event P happened, but rather why event P happened instead of some event Q"* [22].

Causal explanation happens through several processes [10]. First, there is information collection: a person gathers the information available. Second, a causal diagnosis takes place: a person tries to identify a connection between two events/instances based on the information. Third, there is causal selection, a person dignifies a set of conditions as "the explanation". This selection process is influenced by the information gathered and the domain knowledge of a person [20]. This means that what people consider acceptable and understandable is selected from the information provided and depends on people's own domain knowledge or role. According to Lambrozo, explanations that are simpler are judged more likely to be believed and more valuable [18] and another study also highlighted that users prefer a combination of simple and broad explanations [26].

As mentioned previously, explanation can be seen as an act or can be seen as a product. Explanation as an act involves the interaction between one or more explainer and explainee [22]. According to Hilton, explanation is understandable only when it involves explainer and explainee engaging in information exchange through dialogue, visual representation, or other communication modalities [10]. This statement implies that static explanations could be harder to understand because they could be less engaging and would not involve a dynamic interchange between explainer and explainee. To achieve meaningful explanation, a social (interactive) characteristic of explanation needs to be taken into account.

Previous research also showed that participants place the highest trust in explanations that are sound and complete [14]. Soundness here means nothing but the truth, how truthful each element in an explanation is with respect to the underlying system. Completeness here means the whole truth, the extent to which an explanation describes all of the underlying system. Completeness is argued to positively affect user understandability [13]. Even though both of Kulesza's studies used explanation in the case of a music recommender system, we think that being truthful (soundness) and thorough (completeness) are key characteristics of explanations to be further explored. Building on the literature reviewed above, we therefore distilled 6 key characteristics of meaningful explanation, that are defined in Table 1.

Table 1: Characteristics of Meaningful Explanation

Explanation	Description
contrastive	the cause of something relative to some other thing in contrast [16][10][22]
domain/role dependent	pragmatic and relative to the background context [10][20]
general	simpler and broad explanation is preferable [26][18]
social/interactive	people explain to transfer knowledge, thus can be a social exchange [10][22]
truthful	how truthful each elements in an explanation is with respect to the underlying system [14]
thorough	describes all of the underlying system [14]

2.2 Explanation and User's Trust

There is arguably a relation between explanation and users' trust. According to the Defense Advanced Research Projects Agency (DARPA), Explainable AI is essential to enable human users to understand and appropriately trust a machine learning system [8]. Previous studies proposing different types of explanation [27][2][9][14] further cemented the claim that explanations improves user trust [30][24][5].

However, users' trust could be misplaced and lead to over-reliance or over-trust. In a healthcare scenario, a doctor could unknowingly trust a technologically complex laboratory diagnostic test that incorrectly calibrated and misdiagnosed patients [4]. Previous research suggests that giving explanation could help users to moderate their trust level [31], either by providing explanation as system's accuracy [33][23] or as system's confidence level [15]. On one hand, these findings are not applied to healthcare. Hence, while system's accuracy and system's confidence level might be highly affecting users' trust in dating app [33], or context aware app [15], it is unclear if that would be the case in a healthcare scenario. On the other hand, in the healthcare/medical domain, Bussone et al. found that a high system's confidence level had only a slight effect on over-reliance [3].

There are a number of ways to present an explanation. For example, a study mentioned above, used accuracy level as explanation. It is important to know, what kind of style we are going to present our explanation. Research found that explanation style and modalities affect users trust toward algorithmic systems, with the result that this can either improve or decrease [12][25].

In addition, in each of the reviewed studies trust was measured differently, hence the results are hard to compare and do not provide

a clear picture of the extent to which different styles of explanation affect different types of trust. To better understand users' trust towards an AI medical system, a more comprehensive trust measurement instrument is needed and will be explored in the next section.

2.3 Trust Measurement

In general, there is quite a large literature presenting scales for measuring trust. This paper will focus on identifying an appropriate scale for the assessment of human trust in a machine prediction system, which can be contextualised to a healthcare scenario.

Some of the trust measurements reviewed from the automation literature are highly specific to particular application contexts. For example, the scale developed by Schaefer [28] refers specifically to the context of human reliance on a robot. The questions that are asked to users to measure trust are, for example: "Does it act as part of a team?" and "Is it friendly?". Another example of specific trust measurement is the scale developed by Dzindolet, et al. [6]. It was created in the context of aerial terrain photography, showing images to detect camouflaged soldiers. The questions asked to measure trust in this case are for example: "How many errors do you think you will make during the 200 trials?". As these questions are very specific to the task and the technical knowledge of the users in the specific application context, it would be hard to translate them to a healthcare scenario.

Madsen and Gregor [19] developed and tested a more generic human-computer trust measurement instrument, with the focus on trust in an intelligent decision aid. A validity analysis conducted of this instrument showed high Cronbach's alpha results, which makes this scale promising to be tested in a different application field. Trust factors here are divided in two groups, cognitive based trust and affect based trust. Madsen and Gregor [19] conceptualise trust as consisting of five main factors: perceived reliability, perceived technical competence, perceived understandability, faith, and personal attachment. Perceived Technical competence means that the system is perceived to perform the tasks accurately and correctly, based on the input information. Perceived Understandability means that the user can form a mental model and predict future system behaviours. Perceived Reliability means that the system is perceived to be consistently functioning. Faith means that the user is confident in the future ability of the system to perform, even in situations in which has never used the system before. Finally, personal attachment means that users find using the system agreeable, preferable, and that suits their personal taste.

Some of these factors overlap with the trust factors identified by McKnight [21]. McKnight provides an understanding of trust in technology in a wider societal context. McKnight [21] defines trust as consisting of three main components: propensity to trust general technology, institution-based trust in technology, and trust in specific technology. In the context of this paper we only focus on trust in a specific technology. McKnight [21] defines trust in a specific technology as a person's relationship with a particular technology. Even if the study does not specifically target decision systems, the paper goes into a large literature and looks at different object of trust, trust attributes, and their empirical relationships, thus proposing a scale of trust which demonstrated good reliability with Cronbach's alpha > 0.89. In the proposed scale, trust with a specific technology was analyzed into three factors: perceived functionality, perceived helpfulness, and perceived reliability. Perceived functionality is users' perceived capability of the system to properly accomplish its main function. Perceived helpfulness is users' perception of the technology providing adequate, effective, and responsive help. Finally, perceived reliability means that the system is perceived to operate continually or responding predictably to inputs.

Table 2: Human-AI Trust Measurement

Trust Factors	Description
perceived technical ability	system is perceived to perform the tasks accurately and correctly based on the information that is input.
perceived reliability	system is perceived to be, in the usual sense of repeated, consistent functioning.
perceived understandability	user can form a mental model and predict future system behaviour.
personal attachment	user finds using the system agreeable, preferable, suits their personal taste.
faith	user has faith in the future ability of the system to perform even in situations in which it is untried.
perceived helpfulness	user beliefs that the technology provides adequate, effective, and responsive help.

In our study we adopt a merged and modified version of the 9 trust items proposed by Madsen and Gregor and by McKnight. From the total 9 trust items, that have been described above, we merged items that overlapped in meaning and modified some of their descriptions into the final 6 trust metrics: perceived understandability, perceived reliability, perceived technical competence, faith, personal attachment, and helpfulness (See Table 2).

3 METHODOLOGY

We aimed to test to what extent different types of textual explanations affect different factors of users' trust. In section 2.1 we have identified 6 characteristics of meaningful explanation: contrastive, truthful, general, thorough, social/interactive, and role/domain-dependent explanations (see Table 1). We used these characteristics to design distinctive textual explanations, and then presented them to users. Since we focus on a healthcare scenario, we used a dramatising vignette to probe participants responses. We asked them to read the explanation after reading the vignette and then run an online survey asking them to rate different explanation types. To elicit feedback on the explanation types we used the trust measurement mentioned above.

We designed a between-subjects study, in which different groups of users were each presented with a different explanation type. When designing the explanations, we focused on 4 out of the 6 explanation characteristics: contrastive, general, truthful, and thorough. Social/interactive and role/domain-dependent characteristics

Table 3: Explanation Styles

Characteristic	Presented Explanation
contrastive	"From the screen image, Malignant lesions are present. Benign cases and fluid cyst looks hollow and have a round shape. Your spots are not hollow and and have irregular shapes. Therefore, your spots are detected as Malignant."
general	"Based on your screen image, your spots are detected as Malignant. 19 in 20 similar images are in Malignant class."
truthful	"Using 5,600 of ultrasound images in our database, your image have 95% similarities with Malignant cases."
thorough	"Malignant lesions are present at 2 sites, 30mm and 5mm. Non homogeneous. Non parallel. Not circumscribed. Your risk of breast cancer as; 30-50 years old, cyst history, woman is increased 20%"

were ignored at this stage for simplicity. In fact, these explanation styles could not be expressed with a textual description, and needed work on the UX design of the explanation type in order to be realized. Therefore the assessment of the effects of these two characteristics was left for future study. The AI system's diagnosis tool described in the dramatising vignette was a fictional AI system for mammography diagnosis, used in a self managed health scenario. With the system users could upload images of self-scanned mammograms and then received a diagnosis result with an attached textual explanation.

3.1 Explanation Design

In order to design the explanation, we first tried to look at breast cancer diagnosis report and several screening reports including ultrasound. Next, we designed the possible textual explanations based on each characteristic definition in a small-scale informal design phase. We then consulted the designed explanations with researcher outside this study and medical professional. The explanations were identical from a UI perspective, with one graphic and followed by the diagnosis and the explanation text. The explanation texts were designed to stress the four explanation characteristics: contrastive, truthful, general, thorough. We also tried to present a balanced level of system's capability, for example in general style: *"19 in 20 similar images"* and in truthful style: *"95% similarities"*. The explanation text presented to the participants can be seen in Table 3 and how we presented it can be seen in Figure 1.

3.2 Data Collection and Analysis

The participants were recruited on Mechanical Turk, with a survey set up using Google Form. Our target was initially 80 participants, with 40 participants from the general public and 40 participants from worker in the healthcare field. We choose the option of "master worker" and added one check-in question in the survey, to maximise participation quality and check if the participant read the vignette carefully. The Mechanical Turk hits were up for a week, and in the end, we got 48 participants (only 8 with some medical expertise).

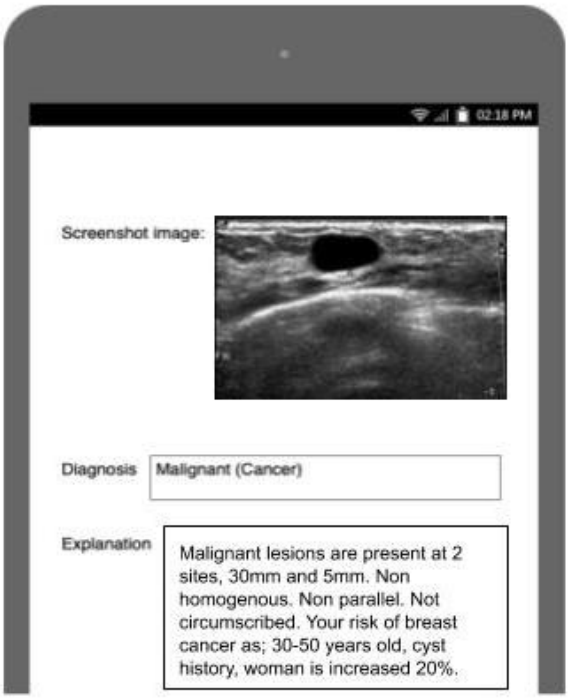

Figure 1: Thorough Explanation

Participants were randomly assigned to 1 of the 4 conditions, with each condition being a different explanation type. The number of participants for each condition are not identical, with $n_1 = 12$, $n_2 = 12$, $n_3 = 11$, and $n_4 = 13$.

We asked participants to rate the AI system after having read the dramatizing vignette and to reflect on the 6 trust's components while rating the explanation using a 7-points Likert scale. Following a between-subject comparison of the results we were able to identify which explanation (if any) affects which of the 6 components of trust, and to what extent. The overall aim of the study was to give us insights on how different styles of linguistics explanations affect specific aspects of users' trust. We also asked participants if they would have liked the presented explanation to be included in the AI system and explain why.

To analyse the data, we used ANOVA tests, followed by Tukey's posthoc paired tests, to see the relative effects of different explanation types. The ANOVA test tells us whether there is an overall difference between the groups, but it does not indicate which specific groups differed. The Tukey's post-hoc tests can confirm where the difference occurred between specific groups. In addition, we evaluated the trust measurement instrument, by using Cronbach's Alpha.

4 RESULTS

From the online survey data, we ran two ANOVA tests, to check the explanation styles and the trust factors. In the first ANOVA test, we compared the 4 explanations types in relation to an average trust factor (calculated as median value between the 6 trust scores).

We found that different styles of explanation significantly affect average trust values (pvalue=0.0033, α=0.05). We then ran a Tukey's posthoc test, and found that general explanation show significantly lower trust scores compared to the rest of the explanation styles; contrastive, truthful, and thorough (α=0.05). The Tukey's posthoc test analysis can be seen in Fig 2.

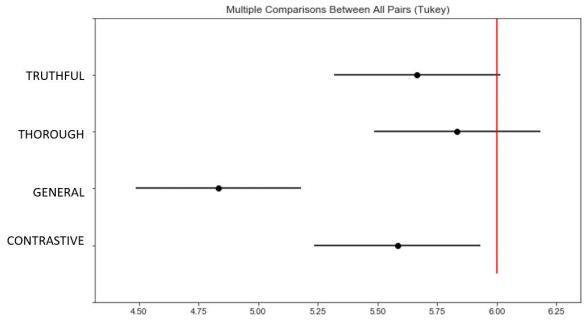

Figure 2: Tukey's post hoc test in explanation styles

In the second ANOVA test, we compared the four explanation styles for each trust factor, we therefore ran 6 comparisons and found that Personal Attachment was the only trust factor showing significant difference (pvalue=0.02158, α=0.05). We then ran a Tukey's posthoc test for Personal Attachment, to identify where the specific difference occurred, and found that contrastive and thorough explanation styles shows significant difference compared to general explanation style (α=0.05). The Tukey's posthoc test analysis can be seen in Fig 3.

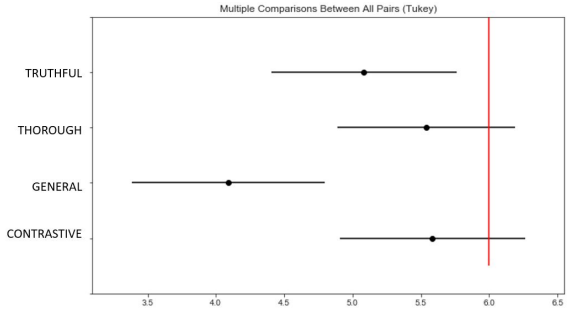

Figure 3: Tukey's post hoc test in Personal Attachment

As mentioned above, other than trust scaling, we also asked participants if they would like the explanation style presented to them to be included in the app for self managed health. We can see in Fig 4, contrastive, truthful, and thorough explanation styles are rated quite high (6 = very), while the general explanation style is rated lower (5 = moderately). This assessment is consistent with the explanation style-trust analysis we did. In the analysis, it shows that general explanation is the least performing explanation in affecting personal attachment.

We also asked why participants preferred or not to receive the explanation given to them. By qualitative analysing the 25 answers from thorough and contrasting style groups, users reported the

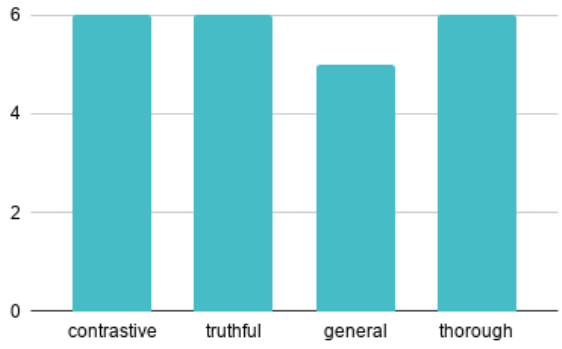

Figure 4: Median of participant's rating towards their explanation preference

presence of a clear rationale, and the use of lay terms, as the two distinctive factors motivating the high trust rating. In turn, the need of a rationale for the AI result was also explicitly mentioned as a way to improve general explanation (by 4 out of 11 people in the general explanation style group mentioned rationale as a need).

The trust measurement was tested using the overall data from 48 participants. The reliability of the overall measurement was determined by Cronbach's Aplha. We found that the alpha is quite high, α=0.88. This is an encouraging result which may inform further use, testing and validation of the proposed human-AI trust measure in other healthcare applications.

5 DISCUSSION

Our study confirms previous research indicating that different styles of explanation significantly affect specific trust factors. In particular we found that Personal Attachment (pvalue=0.02158) was significantly affected by different textual explanation styles, and was highly rated by the groups that were presented with thorough and contrastive explanation styles. This means that among the participant, thorough and contrastive styles suited their taste more, compared to the general explanation style.

This finding was corroborated by the additional comparison of the 4 explanations by average trust ratings, which showed that general style explanation was significantly rated lower than the rest of the explanation styles. Overall preferability scores also confirmed that general style explanation was rated the lowest.

Participants seemed to prefer thorough and contrastive styles explanation because of the rationale provided, and because of the layperson language used to provide the explanation. The need of rationale was also suggested as a way to improve general explanation style.

However, further investigations about the extent to which explanation affects trust judgement need to be conducted. The current results are not conclusive and sufficient to develop an explanation style and trust relation model. Additional studies to explore the explanation mediums and interaction types are also necessary.

6 LIMITATIONS AND FUTURE WORK

This preliminary study has several limitations that should be noted. This is an exploratory study of quite a broad topic and we only conducted one online survey with low number of participants. The fact that some explanation styles did not show significantly different effects on users trust judgements could be caused by the small sample size. Future studies with a bigger sample size and a baseline group are needed to determine the extent of which explanation affects trust.

We also acknowledge that trust is difficult to measure. Even though our trust measurement has shown high internal consistency, we have not fully investigated the validity of the measurement in other cases/fields. Moreover, in this experiments, we only measured user's trust as a self reported measure. Our experimental design, and the use of a probing method, may have also possibly influenced participants' reflection and self reporting. Further research is needed to carefully determine whether this was the case.

REFERENCES

[1] Peter Achinstein. 1983. *The nature of explanation*. Oxford University Press on Demand.

[2] Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. Towards improving trust in context-aware systems by displaying system confidence. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*. ACM, 9–14.

[3] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.

[4] Pat Croskerry. 2009. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Advances in health sciences education* 14, 1 (2009), 27–35.

[5] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

[6] 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.

[7] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820* (2018).

[8] David Gunning. 2017. Explainable artificial intelligence (xai). (2017).

[9] 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*. ACM, 241–250.

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

[11] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* (2017).

[12] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2390–2395.

[13] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1–10.

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

[15] Brian Y Lim and Anind K Dey. 2011. Design of an intelligible mobile context-aware application. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services*. ACM, 157–166.

[16] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements* 27 (1990), 247–266.

[17] Zachary C Lipton. 2017. The Doctor Just Won't Accept That! *arXiv preprint arXiv:1711.08037* (2017).

[18] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.

[19] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

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

[21] D Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)* 2, 2 (2011), 12.

[22] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).

[23] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019).

[24] Alun Preece. 2018. Asking 'Why'in AI: Explainability of intelligent systems–perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management* 25, 2 (2018), 63–72.

[25] Pearl Pu and Li Chen. 2006. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, 93–100.

[26] Stephen J Read and Amy Marcus-Newhall. 1993. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology* 65, 3 (1993), 429.

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386* (2016).

[28] Kristin Schaefer. 2013. The perception and measurement of human-robot trust. (2013).

[29] Saravanan Thirumuruganathan, Mahashweta Das, Shrikant Desai, Sihem Amer-Yahia, Gautam Das, and Cong Yu. 2012. Maprat: Meaningful explanation, interactive exploration and geo-visualization of collaborative ratings. *Proceedings of the VLDB Endowment* 5, 12 (2012), 1986–1989.

[30] Eric S Vorm. 2018. Assessing Demand for Transparency in Intelligent Systems Using Machine Learning. In *2018 Innovations in Intelligent Systems and Applications (INISTA)*. IEEE, 1–7.

[31] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 601.

[32] Sam Wilkinson. 2014. Levels and kinds of explanation: lessons from neuropsychiatry. *Frontiers in psychology* 5 (2014), 373.

[33] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 279.