Trust and Reliance in Consensus-Based Explanations from an Anti-Misinformation Agent

Takane Ueno
ueno.t.ao@m.titech.ac.jp
Tokyo Institute of Technology
Meguro City, Tokyo, Japan

Yeongdae Kim
kim.y.ah@m.titech.ac.jp
Tokyo Institute of Technology
Meguro City, Tokyo, Japan

Hiroki Oura
houra@rs.tus.ac.jp
Tokyo University of Science
Shinjuku City, Tokyo, Japan

Katie Seaborn
seaborn.k.aa@m.titech.ac.jp
Tokyo Institute of Technology
Meguro City, Tokyo, Japan

ABSTRACT

The illusion of consensus occurs when people believe there is consensus across multiple sources, but the sources are the same and thus there is no "true" consensus. We explore this phenomenon in the context of an AI-based intelligent agent designed to augment metacognition on social media. Misinformation, especially on platforms like Twitter, is a global problem for which there is currently no good solution. As an explainable AI (XAI) system, the agent provides explanations for its decisions on the misinformed nature of social media content. In this late-breaking study, we explored the roles of trust (attitude) and reliance (behaviour) as key elements of XAI user experience (UX) and whether these influenced the illusion of consensus. Findings show no effect of trust, but an effect of reliance on consensus-based explanations. This work may guide the design of anti-misinformation systems that use XAI, especially the user-centred design of explanations.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**; • **Information systems** → Social networks.

KEYWORDS

intelligent agent, explainable AI, consensus, misinformation, user experience, trust, reliance

ACM Reference Format:
Takane Ueno, Yeongdae Kim, Hiroki Oura, and Katie Seaborn. 2023. Trust and Reliance in Consensus-Based Explanations from an Anti-Misinformation Agent. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3544549.3585713

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585713

1 INTRODUCTION

AI is increasingly being taken up for a variety of purposes, thanks to complex algorithms and the availability of big data. At the same time, a user-centred problem with AI is coming into sharper relief: that the algorithms making up AI systems are usually black boxes that make decisions in ways not transparent to people. Even when they are transparent, people are often unable to understand because the output is too complex. Indeed, prediction accuracy and explainability in AI are generally trade-offs [27]. To address these issues, the notion of explainable AI (XAI) has gained momentum in recent years. The purpose of XAI, a term coined by DARPA researchers in 2017, is to make AI decision-making and other fundamental behavior more understandable to people by providing a human-parsable explanations [8].

Many of us, members of the general public, are starting to use these seems (or desire to). For example, many people have become concerned with the level of misinformation on social media and desire intelligent fact-checking tools [31], notably on Twitter, and perhaps especially now after its acquisition by Elon Musk and subsequent revocation of fail-safe measures [29]. Still, few XAI initiatives and specifically the forms of the explanations provided are designed with general non-professional end-users in mind, i.e., the lay public. A recent survey found that machine learning (ML) model descriptions are primarily used by ML engineers to debug models during the development phase [1]. Furthermore, most work appears to rely on the researchers' own intuitions, i.e., expertise, of what constitutes a good explanation [22].

A pressing question is what forms of explanations are effective at enabling lay end-users to trust XAI systems [2, 30, 33, 34]. In social media platforms, lay people make judgments and decisions based on evaluating and integrating reports from multiple informants [4, 31]. One factor that affects lay people's confidence in such reports is the degree of consensus across related reports [4]. In social media, people may desire explanations that rely on consensus-based data sources. Placing these sources within explanations provided by anti-misinformation XAI may increase users' confidence in the system's fact-checking and possibly reliance on it. Yet, this poses a new problem: an *illusion of consensus* effect [4, 36], where people are unable to distinguish "true" consensus, i.e., different informants relying on different sources but drawing the same conclusion, and "false" consensus, i.e., different informants relying on the *same* source.

This raises three questions at the intersection of user-centred design and research methods. First, are people able to distinguish true and false consensus, i.e., not fall prey to an illusion of consensus, when explanations are provided by a user-centred XAI system? Second, how do we distinguish those who are simply skeptical of AI? And third, if the illusion occurs, can it be decreased by emphasizing the independence of sources via the design of the explanations, as suggested by work outside of social media [4]? To the best of our knowledge, this has not yet been explored. Therefore, we asked: *How does consensus relate to user trust and reliance on the use of an intelligent anti-misinformation XAI agent?* Specifically: *(RQ1) If the agent provides a consensus-based explanation about its fact-checking, does this lead to increased trust (attitude) and/or reliance on the agent (behavior)?* We also asked: *(RQ2): If an illusion of consensus appears, what effect, if any, does the consensus-based explanation have on trust and/or reliance?*. To this end, we conducted a comparative evaluation of a prototypical XAI agent within a live Twitter environment that provided explanations to lay Twitter users about its fact-checking decisions. The main contributions of this work are: (i) initial empirical attitudinal, behavioural, and user experience (UX) evidence of a relationship between reliance, but not trust, on fact-checking services provided by a consensus-based XAI agent and subsequently (ii) evidence of an illusion of consensus effect. This work highlights the importance of consensus and its presentation in XAI systems using the case study of misinformation on Twitter.

2 BACKGROUND

2.1 Trust and Reliance in AI

Trust in automation is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [18, p. 54]. This definition is based on a strong foundation of empirical research and frequently referenced in the context of AI [30]. Lee and See [18] define trust as an *attitude* and distinguish it from *trust* as a behavior. Similarly, Hoff and Bashir's [11] model of trust in automation makes a clear distinction between trust as an attitude and behavior, characterizing trust as a factor that mediates automation performance and user behavior, or *reliance*. Most recently, Papenmeier, Kern, Englebienne and Seifert [25] reported discrepancies between self-reported trust and trust as a behavior, indicating that it is important to clearly distinguish between attitude and behavior. In this work, we follow suit by operationalizing trust as an attitude and reliance as a behavior.

Still, these concepts were developed for automation. Indeed, trust and reliance have not been well distinguished in research on AI-based systems and have been grouped together under the term trust [30]. Trust is typically measured in isolation as a subjective attitude through questionnaires and interviews [12, 16, 34]. In some studies, trust has been approached as *dependent* behaviors or biological responses [23, 32]. Other research has considered the roles of automation performance and dependence as mediators of trust [3, 21, 35]. The focus has been on the relationship between automation reliability, dependence, and trust, and the results of these studies are somewhat contradictory. Hussein, Elsawah and Abbass [13] reexamined this literature and developed experimental guidelines to reduce errors. They analyzed the role of trust mediation on perceived reliability of and dependence on a target sensing system in a flight task. In order to clearly distinguish between trust and reliance, we applied their guidelines to explore the relationship between these factors in the context of consensus-based explanations provided by XAI.

2.2 Appraising Consensus Across Reports and Sources

In daily life, we rely on consensus when evaluating and integrating various pieces of information to make decisions [4]. However, information is not always independent; separate pieces of information may use the same source/s. For example, over 80% of climate change denial blogs relied on a single primary source [9]. Reliance by multiple independent informants on a single source of data is called a *false consensus* and can influence the formation of accurate beliefs. Yousif, Aboody and Keil [36] investigated perceptions false consensus. Subjects were were assigned to one of a true consensus condition in which they read four positive sentences with different primary sources and one sentence with a negative primary source, a false consensus condition in which they read four positive sentences with a single primary source and one sentence with a negative primary source, a false consensus condition in which they read a positive sentence and a negative, and a baseline condition in which they read one sentence each, and after reading each sentence, they were asked how much they agreed with the assertion. As a result, they discovered an *illusion of consensus*, in which subjects gave similar agreement ratings to presentations of true and false consensus. Connor Desai, Xie, and Hayes [4] investigated this consensus illusion, believing that its creation was due to people's perception of the independence of information sources. They followed the same experimental procedure as Yousif et al. [36] but also highlighted single sources of information with the same color, emphasizing the relationship between each source. As a result, true consensus, with its emphasis on independence, received greater agreement than false consensus. They further investigated this in the context of an election poll on Twitter and showed that people assigned more epistemic weight to true consensus than to false consensus when the relationship between sources was made transparent. Consensus thus appears to affect our level of agreement with an opinion. But do these findings translate to consensus-based explanations provided by XAI? Depending on the type of consensus provided, people may assign different levels of agreement to XAI explanations, thus mediating their trust and reliance on the XAI. Also, if these explanations are transparent about the relationship between data sources, the illusion of consensus may not occur, even when the agent provides the consensus. We follow Hussein et al. [13]'s experimental design by extending the theory of Yousif et al. [36] and Connor Desai et al. [4] to explore the XAI context and these possible effects in this work.

3 THEORETICAL FRAMEWORK

Given the lack of established models of trust in AI and particularly XAI [30], we used Hoff and Bashir's model for automation [11, 13]. Our instantiation of the model for our XAI agent and construction of hypotheses is in Figure 1.

Agreement with information when there is true consensus tends to be higher than when there is false or no consensus [4, 36]. Still, if

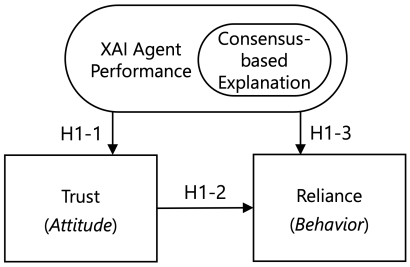

Figure 1: Our theoretical framework, based on Hoff and Bashir's model of trust for automation [11]. Consensus-based explanations provided by the XAI agent influence reliance (behavior), which is mediated by trust (attitude) in the agent's performance.

the trust model for automation [11, 13] applies to XAI systems, then we would expect reliance behavior to be mediated by trust. We thus hypothesize: *(H1-1) "True" consensus-based explanations from agent will increase user trust in agent compared to "false" consensus and "no" consensus.* This effect may also apply to the agent-provided explanations and be reflected in user reliance on the agent, leading to this hypothesis: *(H1-2) User trust in the agent increases user reliance on the agent.* Subsequently, if the results hold true for "true" consensus, and the theorized relationship between trust and reliance exits for XAI systems, then we can also hypothesize: *(H1-3) "True" consensus-based explanations from the agent will increase user reliance on the agent compared to "false" consensus and "no" consensus.*

Following previous work [4], our XAI agent explicitly *labels* sources of information and the data used by these sources to clarify the relationship between individual sources and data. In other words, these labels of sources and data should clearly indicate to users that each source is independent in a true consensus conditions and not independent in a false consensus conditions. Emphasizing the independence between data across sources reduces the illusion of consensus [4]. Thus, the design of the XAI agent's explanations should prevent the illusion of consensus from occurring. We hypothesize: *(H2) The illusion of consensus will not appear when the source information about the data is explicit.*

4 METHOD

We conducted a within-subjects experiment based on Hussein et al. [13]. We used an intelligent XAI agent designed to support metacognitive behaviors in the face of misinformation on Twitter, specifically related to the COVID-19 pandemic. Our protocol was registered in advance of data collection on July 7th, 2022[1]. We obtained ethics from our IRB.

4.1 Participants

A total of 35 participants (22 men, 13 women, none who identified as another gender) who were fluent in Japanese and used Twitter were recruited. The sample size was determined based on the previous study by Hussein et al. [13]. Participants were recruited from Jikken-baito, a Japanese experiment recruiting website through multiple social media platforms and connection between researchers[2] or directly by the authors.

4.2 System Design

We used a novel Twitter-based intelligent XAI agent called Elemi [15]. The agent, which requires curated content, simulates fact-checking within tweet content, providing links to other tweets and sources. If a tweet contains misinformation, the agent adds a banner to the top-right side of the tweet containing an explanation with tweets and their data sources as references for why the agent regards the tweet as misinformed. As a simulated agent that uses curated content, its accuracy is 100%. Tweets and sources were manually collected and verified by the authors. There were three consensus schemes: *True* (three tweets referring to three different sources), *False* (three different tweets referring to the same source), and *None* (only one tweet). Refer to Figure 2 for an illustration of the agent in action.

Figure 2: The agent creates a banner on the right side of the misinformed tweet to display an explanation with references to tweets and other information sources and data. These were not restricted to Twitter alone and included external sources. Here, a "false" consensus of three different tweets referring to the same source is shown.

4.3 Stimuli

Participants viewed a controlled Twitter timeline with tweets sourced from the COVID-19 hashtag in September 2022: 12 factual and 12 misinformed. The factualness of the tweets was verified by the first author based on at least two different sources. Since we used live Twitter, it was possible for participants to have come across these tweets before. Factual and false tweets were randomly ordered into pairs. No other tweets were in the timeline.

4.4 Procedure

All participants gave informed consent. The experiment was divided into two sessions. In Session 1, participants verbally answered how correct they thought each tweet was on a scale of 0-100 (0: completely wrong, 100: completely right) while viewing the timeline *without* the agent. Then there was a 5-minute break.

In Session 2, participants carried out the same procedure. However, this time false tweets were pointed out by the agent. The agent's three consensus conditions ("true", "false", "no") were counterbalanced and changed every eight tweets to account for individual differences. Note that, due to the Musk acquisition of Twitter, some tweets which the agent serve as data sources unexpectedly became unavailable. To mitigate the impact on trust and reliance,

[1] https://osf.io/s7wqe

[2] https://www.jikken-baito.com

Table 1: Results for the random intercept model. β_1-β_3 represent the coefficient of the no consensus dummy variable on true consensus, the coefficient of the false consensus dummy variable on true consensus, and the coefficient of trust, respectively. The marginal and conditional coefficients of determination R^2_m and R^2_c and the Akaike Information Criterion (AIC) were computed. *: $p < .05$

Response Variables	Model	β_1	β_2	β_3	W-S Variance	B-S Variance	R^2_m	R^2_c	AIC
Trust	M_0	-0.05 [-0.22, 0.12]	-0.05 [-0.23, 0.12]	-	0.36	0.66	0.001	0.65	638.6
Reliance	M_0	-0.22 [-0.45, 0.01]	−0.33* [-0.56: -0.10]	-	0.66	0.35	0.02	0.36	775.1
	M_1	-0.22 [-0.45, 0.01]	−0.33* [-0.56, -0.09]	0.023 [-0.11, 0.16]	0.67	0.36	0.02	0.36	780.6

participants who witnessed tweets being unavailable were asked to imagine the existence of other tweets similar to those provided by the agent. To accommodate the dynamic nature of trust [13], participants paused after reading two tweets (one factual and one false) and completed a questionnaire with the trust measures (4.5.1) on a separate tablet.

After the sessions, participants completed a post-experiment questionnaire that included demographics and open-ended questions: "How did you feel overall about your experience with the agent (Session 2)? Why did you feel that way? Please be specific." "What did you feel about the data that the agent used to identify tweets with potentially false content? Why did you feel that way? Please be specific." Participants were then thanked and compensated.

4.5 Measures

All measures were translated into Japanese by the authors and back-translated using DeepL, checked by those fluent in both languages. All references to "the system" in the instruments were changed to the agent's name.

4.5.1 Trust. We used the Trust in Automation scale [14], a 7-point Likert scale consisting of 12 items: 7 for trust and 5 for distrust. Although developed for automation, it is also the most commonly used measure for AI [30]. Trust and distrust can exist simultaneously and are different concepts [19, 28]. In our case, we measured trust multiple times (refer to 4.4). In consideration of participant time and workload, only the seven items related to trust were used.

4.5.2 Reliance. Reliance was measured using Weights of Advice (WOA) [10]. While WOA has been used in AI and XAI research as a measure of trust [20, 24, 26], we used it as a measure of reliance because it is an objective indicator of behavior rather than subjective and attitudinal. WOA quantifies the extent to which participants change their ratings as a result of an informant's advice: $WOA_{ij} = (F_{ij} - I_{ij})/(A_{ij} - Iij)$, where I, F, and A denote the initial estimate, the final estimate, and the advisor's advice for some participant i on some trial j, respectively. A WOA of 1 indicates adoption of the advice, 0 indicates maintenance of the initial estimate, and between 0 and 1 indicates that the advice is partially discounted. Notably, a WOA of 0.5 indicates equal weighting of one's own estimate and the advisor's advice. The values of Session 1 and Session 2 were assigned to I and F, respectively. The agent gave advice only on tweets that contained false content, so the WOA was computed only for tweets containing false content and was fixed at A = 0. The value of WOA was truncated to 0 for values less than 0 and to 1 for values greater than 1, following previous studies [5, 6, 20]. Note that some previous studies used absolute values when calculating WOA measurements. For robustness, we also used the absolute value approach, but the nature and significance of the results remain the same.

4.6 Data Analysis

Data were measured by subject for each condition. One person's data was excluded because they rated all measures at 50%, indicating an inability to make judgments about correctness. Data in columns with initial estimate (I) equal to the advice (A) were excluded according to previous work [5, 6]. In the end, 289/315 points of data were analyzed. We fitted random intercept models; all models contain subject as a random effect and consensus (and possibly trust) as fixed effects. We also averaged the results per consensus condition and ran one-way repeated measures ANOVAs.

An applied thematic analysis [7] was conducted on the open-ended responses to explore factors that may influence trust and reliance and compare awareness of the consensus and number of sources. A lead rater developed the initial themes, and then two raters coded all data separately. Inter-rater reliability was assessed by Cohen's kappa [17] with 0.7+ as the criterion for agreement. Themes that did not meet this criterion were modified, merged, or discarded and repeated until the kappa exceeded 0.7. For coding that did not match, disagreements were resolved by discussion.

5 RESULTS AND DISCUSSION

Figure 3 and Table 1 summarize the results of the statistical analyses. Table 2 shows the thematic framework.

We begin with the linear mixed model. The Cronbach's alpha for trust was α = 0.90 and the intraclass correlation coefficient (ICC) for trust was 0.65. In the trust model M_0, consensus is incorporated as a dummy variable based on the true condition because of the categorical three conditions. Neither coefficient was statistically significant ($p > .05$) The ICC for reliance was 0.34. In the reliance model M_0, consensus is incorporated, with β_1 not statistically significant ($p > .05$) but β_2 = -0.33, 95% confidence interval (CI) = [-0.56,

Table 2: Thematic analysis framework. Numbers in parentheses are frequencies for Q1 and Q2, respectively.

Theme	Sub-theme	Definition	Examples
The agent and its algorithm	Usefulness of the agent (29, 5)	The agent's "fact-checking" and sources were useful and effective, or not.	"It was useful to know which tweets might include false and the tweets."
	Concerns about misuse of the agent (10, 2)	Fear of or actual over-reliance and misuse of the agent.	"The agent almost made me decide that the information was false without checking the link."
	Consistent with the agent (4, 2)	Agent and subject agreed on factualness of the content, or not.	"My own feeling matched the agent's many times, "
	Time lapse (3, 2)	Subjects' trust and reliance on the agent changed over time.	"I was skeptical in the beginning, but I trusted it in the end …"
	Questions about the agent's algorithm (2, 1)	Subjects wanted to understand the agent's algorithm and the criteria by which it made decisions.	"I wondered how they find users who have opposing views."
Tweets/ sources provided by the agent	Reliability of content (3, 20)	Feelings about the reliability of the tweets and sources provided by the agent.	"The agent's data sources also seemed like sites I wasn't sure I could trust."
	Number/ consensus of sources (5, 8)	Trust and reliance on the agent varied based on the number of sources and type of consensus the agent provided.	"I felt that it would be more credible if the sources were not all the same."
	Unavailable tweets (1, 7)	When the agent provided unavailable tweets, subjects felt a loss of trust and reliance on the agent.	"…I have a sneaking suspicion that the fact that it was erased may have been false information in itself…"
	Unfamiliar foreign content (2, 6)	Difficult to judge the agents' support because the tweets and sources provided were foreign.	"I was not familiar with many of the foreign references."
	Heavy use of the same sources (1, 2)	Whether the agent was dependable in providing the same sources repeatedly.	"The fact that the source data were from the same organization made the judgement seem a bit untrustworthy."

-0.10], which was statistically significant (p = .006). Reliance model M_1 was added trust to and was statistically significant with only β_2 = -0.33, p = (.006), 95% CI = [-0.56, -0.093].

An ANOVA indicated a statistically significant difference across consensus conditions, $F(2, 66) = 4.46$, p = .015. A post-hoc analysis with a Bonferroni correction revealed that the pairwise differences between false and true were statistically significantly different (p = .032). Other pairwise differences were not statistically significantly different ($p > .05$). Also, trust was not statistically significant under different consensus conditions, $F(2, 66) = .10$, p = .90.

Ten qualitative sub-themes were identified and further classified into two major themes (Table 2). The theme "The agent and its algorithm" includes references related to trust, dependence, and reliability of the agent itself and its algorithm. The theme "Tweets/sources provided by the agent" refers to the tweets provided by the agent in its explanations, and references related to trust, dependence, and reliability on the fact-checking articles as sources.

The results partially support (H1-2) but do not support (H1-1) and (H1-3) with respect to (RQ1). True consensus significantly increased reliance on the agent compared to false consensus, but did not do better than no consensus. Also, consensus-based explanations did not affect trust in the agent. The agent consistently received high trust scores ($\tilde{5}$/7), which did not mediate the relationship between

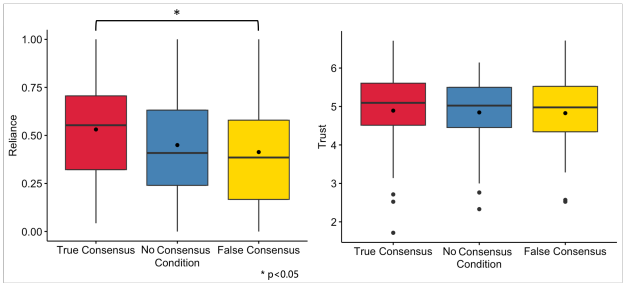

Figure 3: The box-and-whisker diagram shows the results of the ANOVAs. The boxes indicate the interquartile ranges and the horizontal lines are the medians. The circles inside the box represent the mean.

consensus-based explanations and reliance. In short, trust (attitude) towards the agent was not affected by consensus-based explanations. Still, subjects recognized the difference between true and false consensus and relied more on true consensus-based explanations (behavior). Consensus or number of sources was an important factor. Even so, most subjects stated that the function was useful, which may be a factor that sustained such a high level of trust.

Trust does not necessarily manifest itself in reliance, and reliance is not necessarily evidence of trust [22]. Our results reiterate the importance of separating trust and reliance in XAI agents.

The results support (H2) for (RQ2). Subjects relied on the agent's true consensus-based explanations significantly more than those based on false consensus. Significantly, *the illusion of consensus did not appear*. This confirms the results of previous work [4] for XAI explanations: by making the relationships among data sources transparent, our XAI agent prevented an illusion of consensus. Still, unlike in previous work [4, 36], the difference between true and no consensus, i.e., the difference in number of sources used in the explanation, was not significantly different for reliance. The thematic analysis results suggest that subjects focused more on the trustworthiness of individual sources rather than on consensus and the number of data sources. Future work can explore how to raise awareness of this.

This study was limited by its focus on trust and reliance over other influential factors, such as individual differences (e.g. neurotic tendencies), agent reliability, and so on. Incorporating these factors will help us better understand the impact of consensus-based explanations on trust and reliance. Some subjects were unable to use some data source tweet due to Twitter's volatility. Future studies should use research designs or technology hacks that prevent such events or explore their impact. As a lab-based study, our agent relied on a timeline containing 24 tweets. Larger, longer-term studies will be needed to better understand the dynamic nature of trust as well as studies in the wild. Finally, the agent could be explored as a general tool for, e.g., classifying toxicity or predicting risk of posting misinformation.

6 CONCLUSION

Source consensus in explanations affected reliance on the XAI agent, but trust was not a mediator. The illusion of consensus did not occur because the agent ensured that the relationships among the data sources were transparent. Our findings provide initial evidence of the importance of revealing the relationships among data sources in explanations and the importance of providing true consensus in fact-checking XAI agents.

ACKNOWLEDGMENTS

This work was funded by a DLab Challenge: Laboratory for Design of Social Innovation in Global Networks (DLab) Research Grant. We thank Jacqueline Urakami and the Aspire Lab for early design and research feedback.

REFERENCES

[1] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 648–657. https://doi.org/10.1145/3351095.3375624

[2] Andrea Brennen. 2020. What do people really want when they say they want "explainable AI?" We asked 60 stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3383047

[3] Eric T. Chancey, Alexandra Proaps, and James P. Bliss. 2013. The role of trust as a mediator between signaling system reliability and response behaviors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57, 1 (Sept. 2013), 285–289. https://doi.org/10.1177/1541931213571063 Publisher: SAGE Publications Inc.

[4] Saoirse Connor Desai, Belinda Xie, and Brett K. Hayes. 2022. Getting to the source of the illusion of consensus. *Cognition* 223 (June 2022), 105023. https://doi.org/10.1016/j.cognition.2022.105023

[5] Francesca Gino. 2008. Do we listen to advice just because we paid for it? The impact of advice cost on its use. *Organizational Behavior and Human Decision Processes* 107, 2 (Nov. 2008), 234–245. https://doi.org/10.1016/j.obhdp.2008.03.001

[6] Francesca Gino and Don A. Moore. 2007. Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making* 20, 1 (2007), 21–35. https://doi.org/10.1002/bdm.539 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.539.

[7] Greg Guest, Kathleen M. MacQueen, and Emily E. Namey. 2012. *Applied Thematic Analysis*. SAGE Publications, Thousand Oaks, CA, USA.

[8] David Gunning, Eric Vorm, Jennifer Yunyan Wang, and Matt Turek. 2021. DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters* 2, 4 (2021), e61. https://doi.org/10.1002/ail2.61 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.61.

[9] Jeffrey A Harvey, Daphne van den Berg, Jacintha Ellers, Remko Kampen, Thomas W Crowther, Peter Roessingh, Bart Verheggen, Rascha J M Nuijten, Eric Post, Stephan Lewandowsky, Ian Stirling, Meena Balgopal, Steven C Amstrup, and Michael E Mann. 2018. Internet Blogs, Polar Bears, and Climate-Change Denial by Proxy. *BioScience* 68, 4 (April 2018), 281–287. https://doi.org/10.1093/biosci/bix133

[10] Nigel Harvey and Ilan Fischer. 1997. Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. *Organizational Behavior and Human Decision Processes* 70, 2 (May 1997), 117–133. https://doi.org/10.1006/obhd.1997.2697

[11] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (May 2015), 407–434. https://doi.org/10.1177/0018720814547570 Publisher: SAGE Publications Inc.

[12] Marius Hoggenmüller, Martin Tomitsch, Luke Hespanhol, Tram Thi Minh Tran, Stewart Worrall, and Eduardo Nebot. 2021. Context-Based Interface Prototyping: Understanding the Effect of Prototype Representation on User Feedback. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445159

[13] Aya Hussein, Sondoss Elsawah, and Hussein A. Abbass. 2020. Trust Mediating Reliability–Reliance Relationship in Supervisory Control of Human–Swarm Interactions. *Human Factors* 62, 8 (Dec. 2020), 1237–1248. https://doi.org/10.1177/0018720819879273 Publisher: SAGE Publications Inc.

[14] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (March 2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04 Publisher: Routledge _eprint: https://doi.org/10.1207/S15327566IJCE0401_04.

[15] Yeongdae Kim, Takane Ueno, Katie Seaborn, Hiroki Oura, Jacqueline Urakami, and Yuto Sawa. 2023. Exoskeleton for the mind: Exploring strategies against misinformation with a metacognitive agent. In *Proceedings of the 2023 ACM International Conference on Augmented Humans (AHs)*. ACM, Glasgow, Scotland, UK. https://doi.org/10.1145/3582700.3582725

[16] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300717

[17] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. https://doi.org/10.2307/2529310 Publisher: [Wiley, International Biometric Society].

[18] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 Publisher: SAGE Publications Inc.

[19] Roy J. Lewicki, Daniel J. McAllister, and Robert J. Bies. 1998. Trust and Distrust: New Relationships and Realities. *The Academy of Management Review* 23, 3 (1998), 438–458. https://doi.org/10.2307/259288 Publisher: Academy of Management.

[20] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (March 2019), 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[21] Stephanie M. Merritt. 2011. Affective processes in human–automation interactions. *Human Factors* 53, 4 (Aug. 2011), 356–370. https://doi.org/10.1177/0018720811411912 Publisher: SAGE Publications Inc.

[22] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[23] Kazuo Okamura and Seiji Yamada. 2020. Calibrating Trust in Human-Drone Cooperative Navigation. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1274–1279. https://doi.org/10.1109/RO-MAN47096.2020.9223509 ISSN: 1944-9437.

[24] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3491102.3502104

[25] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Transactions on Computer-Human Interaction* 29, 4 (2022), 35:1–35:33. https://doi.org/10.1145/3495013

[26] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–52. https://doi.org/10.1145/3411764.3445315

[27] Arun Rai. 2020. Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science* 48, 1 (Jan. 2020), 137–141. https://doi.org/10.1007/s11747-019-00710-5

[28] Sim B. Sitkin and Nancy L. Roth. 1993. Explaining the Limited Effectiveness of Legalistic "Remedies" for Trust/ Distrust. *Organization Science* 4, 3 (1993), 367–392. https://www.jstor.org/stable/2634950 Publisher: INFORMS.

[29] The Associated Press. 2022. Twitter will no longer enforce its COVID misinformation policy. *NPR* (Nov. 2022). https://www.npr.org/2022/11/29/1139822833/twitter-covid-misinformation-policy-not-enforced

[30] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. 2022. Trust in human-AI interaction: Scoping out models, measures, and methods. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3491101.3519772

[31] Jacqueline Urakami, Yeongdae Kim, Hiroki Oura, and Katie Seaborn. 2022. Finding strategies against misinformation in social media: A qualitative study. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. ACM, New Orleans, LA, 1–7. https://doi.org/10.1145/3491101.3519661

[32] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* 52 (May 2014), 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

[33] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do you trust me?": Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. ACM, Paris France, 7–9. https://doi.org/10.1145/3308532.3329441

[34] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2021. "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces* 15, 2 (June 2021), 87–98. https://doi.org/10.1007/s12193-020-00332-0

[35] Rebecca Wiczorek and Dietrich Manzey. 2010. Is operators' compliance with alarm systems a product of rational consideration? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 54, 19 (Sept. 2010), 1722–1726. https://doi.org/10.1177/154193121005401976 Publisher: SAGE Publications Inc.

[36] Sami R. Yousif, Rosie Aboody, and Frank C. Keil. 2019. The illusion of consensus: A failure to distinguish between true and false consensus. *Psychological Science* 30, 8 (Aug. 2019), 1195–1204. https://doi.org/10.1177/0956797619856844 Publisher: SAGE Publications Inc.