Conversational Agents Trust Calibration

A User-Centred Perspective to Design

MATEUSZ DUBIEL, University of Luxembourg, Luxembourg
SYLVAIN DARONNAT, University of Strathclyde, United Kingdom
LUIS A. LEIVA, University of Luxembourg, Luxembourg

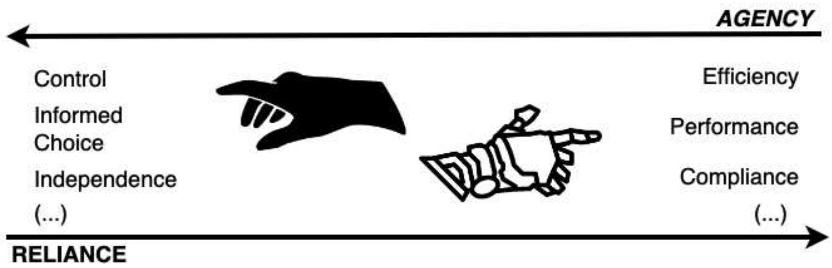

Fig. 1. The spectrum of trust: Illustrates trade-offs between user agency and reliance on conversational agents.

Previous work identified trust as one of the key requirements for adoption and continued use of conversational agents (CAs). Given recent advances in natural language processing and deep learning, it is currently possible to execute simple goal-oriented tasks by using voice. As CAs start to provide a gateway for purchasing products and booking services online, the question of trust and its impact on users' reliance and agency becomes ever-more pertinent. This paper collates trust-related literature and proposes four design suggestions that are illustrated through example conversations. Our goal is to encourage discussion on ethical design practices to develop CAs that are capable of employing trust-calibration techniques that should, when relevant, reduce the user's trust in the agent. We hope that our reflections, based on the synthesis of insights from the fields of human-agent interaction, explainable ai, and information retrieval, can serve as a reminder of the dangers of excessive trust in automation and contribute to more user-centred CA design.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing design and evaluation methods**; *Auditory feedback*.

Additional Key Words and Phrases: Conversational Agents, Design ethics, Trust, User-centred Design

ACM Reference Format:
Mateusz Dubiel, Sylvain Daronnat, and Luis A. Leiva. 2022. Conversational Agents Trust Calibration: A User-Centred Perspective to Design. In *4th Conference on Conversational User Interfaces (CUI 2022), July 26–28, 2022, Glasgow, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543829.3544518

1 INTRODUCTION

Trust has been identified as the key factor influencing users' behaviour during interactions with Conversational Agents (CAs) and a crucial prerequisite for their adoption [38]. Research shows that CAs that sound like humans (e.g. [18, 37]), and look like humans (e.g. [26, 42, 57]) are perceived to be significantly more trustworthy than CAs with more "robot-like" qualities. While the advantages of increasing trust are manifold, including quicker task execution and better performance [19, 40], excessive trust can also have harmful consequences, such as complacency and loss of control over task execution [13, 34]. Following the central premise of Shneiderman's *The Human-Centered Artificial Intelligence* framework: 'people are different from computers' [43, p.502], this paper discusses potential benefits of a non-anthropomorphic CA design as a way to increase user-agency, improve agent's transparency and enable better calibration of trust. Specifically, we consider trust in the context of transactional interactions, where CAs' recommendations can have a consequential impact on the user (e.g. financial implications) as compared to societal/relational interactions that are not goal-oriented by nature (cf. [11]). Recent examples of CA transactional studies include: takeaway-ordering [23], flight-booking [15, 16] and investment-making [47] scenarios.

The contribution of this paper is two-fold. Firstly, through literature review, it bridges insights from the human-agent interaction, explainable AI, and information retrieval disciplines to highlight the dangers of excessive trust and discuss their impact on the user. Secondly, building on these insights, it proposes four design suggestions to promote user agency and improve control during interactions with conversational agents.

2 ASPECTS OF TRUST

The concept of trust has been extensively studied in the context of Human-Automation Interaction (HAI), especially from a human factors and ergonomics perspective [25, 30]. As elicited by the work of Hoff and Bashir [22], trust is an ever-evolving concept influenced by a number of factors (both internal and external) that impact the interaction between a user and an automated system. Some of these factors are linked to cultural and personal biases (e.g. experience of previous interactions with agents), while others are linked to performance of the agent (e.g. accuracy and precision). Recording and measuring trust, as well as its calibration over time, is a highly complex task, which is why trust is often studied in parallel with *reliance* [3] – the degree to which the user depends on the system.

In the context of Information Retrieval (IR), trust is frequently linked to the concept of user *control* [7] and search *agency* [48]. An interactive information search study by White and Ruthven [53] indicated that while users were willing to delegate the task of recommending potential keywords to a search system, they still wanted to retain the control of adding the keywords themselves. In another study, Andolina et al. [5] developed a prototype of a search agent that screened participants conversations for key entities and then used them to proactively retrieve results. Andolina et al. also found that the prototype improved the collaborative search experience by enabling participants to maintain eye-contact. The improvement, however, came at the cost of participants feeling less in control of the search process and, consequently, trusting the agent less. The above findings highlight that proactive CAs might be detrimental to trust and user search agency if the agent has too much control over the search process.

The importance of trust calibration (i.e., adjusting the 'right' level of trust) has also been acknowledged in the literature on CA design [38]. Rheu et al. highlighted that previous research was overly focused on using human-like features to enhance user trust instead of illuminating CAs' features to adjust trust based on the actual capabilities of the system [38]. As stated by Rheu et al., the current CA design practices are based on the implicit assumption that enhancing trust through design features is the ultimate goal. This assumption does not consider the fact that excessive trust could lead to a misguided reliance on the agent which may lead to loss of agency and frustration if user's expectations regarding the CA's capabilities are too high [12]; or under-utilisation of skills if the expectations are too low compared to its actual capabilities (cf. [25, 27]).

As capabilities of CAs rapidly evolve, we postulate that conversational agents should promote trust calibration mechanisms so that users can retain control over the level of their agency and reliance during interaction (see Figure 1), as required for a specific task. In the following sections, we will first comment on the dangers of excessive trust (Section 3) and then provide several design suggestions on how to effectively reduce trust to promote user agency and reduce complacency (Section 4).

3 DANGERS OF EXCESSIVE TRUST

Continued advances in Automatic Speech Recognition [55], Speech Synthesis [20] and Spoken Language Processing [1] are making interactions with CAs increasingly more natural and seamless. While CAs are currently mostly used for simple tasks such as checking weather, factoid queries, playing music or setting alarms [4], a growing number of people is expecting to routinely use them for purchasing products and services [46]. Some of the commercially available CAs already feature online shopping[1] or takeaway-order[2] functionalities, that can directly impact user finances. As the capabilities of CAs improve, it is timely to consider negative implications that miscalibrated trust in such devices can have on users.

Research shows that as familiarity with a system grows, users report increased trust in its capabilities [32]. In the context of conversational user interfaces, familiarity with an agent's voice can lead to the user being less critical of provided recommendations. One of the domains where excessive trust and a lack of user agency can be harmful is online-shopping. A recent study found that an increased facility of purchasing products via 'one-click' button [21] led to an increase in impulsive buying behaviours. It was demonstrated that impulsive buying can lead to over-shopping and, in the long run, can cause shopping addictions [33, 56]. The majority of respondents (69%) to a 2021 survey (N = 81) on purchasing behaviours [21] indicated that convenient shipping was their main reason for frequently using online shopping. Due to the continued COVID-19 emergency, this trend is likely to become even more prevalent in the future.

Another example is the financial sector, where CAs are deployed to help bank customers in selecting loans or mortgages, or to advise investors on how to best manage their assets (e.g. stock transactions). In this context, CAs recommendations are driven by statistical models that are based on historical data and may not generalise well given the dynamic nature of global markets, which could consequently result in users being misguided and taking ill-informed decisions. A study by Ng et al. [31] found that simple socio-emotional features (such as a human-sounding name compared to a more "robotic" one) made participants significantly more likely to share personal banking details and trust the agent.

Some researchers argue that in order to promote better trust calibration, in certain cases it may be desirable to decrease trust (e.g. [14, 38]) or even instil distrust [36], as familiarity developed through repeated interactions with a system may result in forming habits that can have negative consequences for the user. For example, over-trusting the capabilities of a system can form complacent behaviours, often defined as 'an inappropriate monitoring or checking of automated functions' [28]. Complacent behaviour can be harmful to the user in the long term as it increases the risk of missing automation failures and reduces the situational awareness of the user, ultimately resulting in poor

[1]See e.g. https://tinyurl.com/rn7jy6r2 (Last Accessed: 11th April 2022)
[2]See e.g. https://tinyurl.com/4sskb5tv (Last Accessed: 11th April 2022)

decision-making [35]. In an effort to prevent users from over-relying on a system, HAI and XAI (eXplainable AI) studies have focused on increasing transparency of the agent's reasoning [9].

4 CALIBRATING CA TRUST: DESIGN SUGGESTIONS & FUTURE DIRECTIONS

Given the importance of an adequate trust-calibration, we would like to propose four reflection points (i.e. (1) user in control of agent's features, (2) fostering reflections and suggesting alternatives, (3) ethical personalisation, and (4) facilitating scrutinisation of results) on how to effectively manage human-CA trust relationships to improve efficiency of interactions *while* preventing complacency. These reflection points seek to address dangers of excessive trust in CAs, as highlighted in Section 3.

4.1 User in control of agent's features

Research in the field of HAI shows that an adequate calibration of trust can be attained by informing users of the inner-workings of an agent [44]. This is achieved through explanations and visualisations that are provided before, during and/or after the interaction with an agent to increase the user's awareness of the agent's strengths and limitations [24]. Based on the above findings, we posit that an increased understanding of a CA's actions could be achieved through an explicit control over its functions. For example, as presented in Table 1, a user should be able to freely enable and disable proactive recommendations offered by the CA and control the level of provided support as required. This functionality would allow users to retain search agency while *also* enabling them to leverage the benefits of an automated support at any point of the search process by adjusting interaction settings.

Table 1. Controlling the level of search support

		Conversational prompts	Level of support
1.	User:	Hello, I am looking for an Italian restaurant for this evening.	
2.	CA:	Hello Tom, OK, please select your desired level of support.	
3.	User:	Proactive recommendation, please.	
4.	CA:	Based on your previous searches and current location, I recommend 'Gusto di Roma'. It has great reviews and they serve your favourite parmigiana dish. Should I book a table for you and Jen at seven?	*Proactive*
5.	User:	No, let's try something different today. Please switch to manual control and search for Italian restaurants in Luxembourg, Belair district.	
6.	CA:	I have found two restaurants in that area, one in a high and the other in a mid-price range. Would you like to hear more details about any of them?	*Command-and-control*

A similar 'command-and-control' approach is routinely used in most safety critical environments where a system can be toggled on and off by the user (e.g. automatic landing systems in aviation, extensively investigated in trust-focused studies [29]). In the context of CUI interactions, the user could be provided with a menu of features that can be customised. As an example, the CA could present the user with the choice of using explicitly 'synthetic' (e.g. emphasising unnatural, robot-like pitch) or 'natural-sounding' voices. This recommendation is in line with Aylett et al. who suggested a more diverse approach to CA voice design that goes beyond naturalness and human-likeness of synthetic voices [6].

4.2 Fostering reflections and suggesting alternatives

In the context of transaction-based interactions such as online shopping where a CA recommends products and services to the user, several safeguards can be put in place. Firstly, in order to promote more informed choices, the CA can ask the user to provide their motivation for selecting a particular product before completing the purchase. There is evidence that this type of intervention has a strong potential to reduce compulsive buying behaviour during individual online shopping, i.e. without the support of the agent (cf. [21]). Secondly, if the CA detects that the user repeatably engages in similar behaviours (e.g. ordering fast-food takeaways everyday), the CA can encourage them to consider alternatives such as home-cooking or provide purchasing statistics to make them reconsider their decisions. Example interactions with reflective (lines 1 and 2a) and non-reflective (lines 1, 2b-4b) CAs are presented in Table 2.

Table 2. Fostering reflection

		Conversational prompts	Reflection mode
1.	User:	Hello, please order my regular from Burger Palace.	
2a.	CA:	Sure Tom, your order has been placed. It is scheduled to arrive in 10 minutes.	*Off*
2b.	CA:	Tom, it is the fourth time this week that you are ordering fast food. Could you tell me why do you want to order from that place again?	*On*
3b.	User:	I've been very busy this week and don't have any time to cook.	
4b.	CA:	OK, I understand. Just to let you know, there are many restaurants with healthier food options that deliver to your area. Would you like to explore some alternatives?	*On*

While a reflective CA can be perceived as confrontational and inappropriate from the perspective of the user, nonetheless, it also has the potential to reduce complacency and foster a healthier use of technology on a long-term basis by encouraging self-reflection about the use of CAs.

4.3 Ethical personalisation

Most state-of-the-art CAs rely on dynamic, reinforcement-based machine learning methods [8, 41] that provide personalised outcomes based on previous interactions with the user. One of the downsides of such methods is the opacity of the agent's motivations and decisions [39]. In order to increase transparency, the CA should be able to justify why it made certain recommendations, explain its interaction policies and suggest alternative strategies to the user. For example, when making a purchase, instead of selecting the shipping method as 'priority' by default or substituting a missing item with an equivalent one automatically, the CA should explain its interaction policy to the user and allow them to amend it if needed. In order to enable this kind of interaction, we propose a hybrid, user-centred approach where the user's preferences are elicited before their first interaction with the agent (e.g. through survey instruments and tutorials), and where different learning methods are introduced and explained so that the user can develop a more accurate mental model of the CA's motivations for any subsequent interactions. This approach, centred on learning and experimenting with different recommendation policies, offers flexibility and encourages more frequent interactions. An example interaction that features ethical personalisation is provided in Table 3.

Table 3. Ethical personalisation

		Conversational prompts	Ethical personalisation
1.	User:	Hello, please go to CinemaHouse.com and book me one ticket for James Bond at 8pm tonight.	
2a.	CA:	Hello Tom, sure, I have made the reservation. Your card had has been debited for 25$. Enjoy the movie.	*Off*
2b.	CA:	Hello Tom, it looks like you are using their website for the first time. Just to let you know, the company reserves leather seats in the 'delux area' by default. Would you like to customise your preferences?	*On*
3b.	User:	Yes, for any cinema reservations, please always book the cheapest available seat and remove any extras. Never reserve a seat in the first row, unless there are no other options.	
4b.	CA:	OK, I will keep these settings for the future. There is a seat available in the middle fifth row (E8) for 18$. Would you like to book?	*On*

The relationship between the user and CA can evolve, and explicitly eliciting users' consent and clarifying recommendation policies offers a better fit to constantly changing capabilities of CAs.

4.4 Facilitating scrutinisation of results

Most studies centred on the explanation of agents' actions rely on visualisations techniques that summarise the inner-workings of the agent [52, 54]. Voice-only interactions suffer from a broadband limitation (non-persistent nature of speech) as providing too much details to the user would induce cognitive overload and, consequently, reduce the agent's performance and usability. Drawing from insights in both the XAI and IR community, we propose a multi-modal approach to tackle this issue. A study by Alipour et al. [2] experimented with different ways to explain an agent's behaviour; either by displaying textual justifications for the agent's choices or by using heat-maps to highlight areas of interests. As a solution to the constrained broadband of voice-only interactions, a CA could propose to display additional information about its decisions through 'complementary' channels. For example, the CA could choose a nearby desktop computer or the user's smartphone to list potential alternatives and provide the reasoning behind their selection, instead of vocally eliciting each option to the user. Table 4 provides an example of how a CA can assist user in scrutinising the results.

In addition to the visual channel, in the example above, Speech Synthesis Markup Language (SSML) can be used to add vocal emphasis to some results and effectively increase their prominence - this approach has been demonstrated to be effective in the context of voice-only interactions [10]. The selection of the channels and methods of visualisation could be consented and pre-programmed by the user during initial 'familiarisation' process, as described in Section 4.3.

Overall, our four design suggestions advocated a non-anthropo-morphic approach to trust calibration. Nonetheless, we admit that in certain contexts, anthropomorphism such as using expressive vocal features to convey information about CA's attitude could be helpful. For example, in English, breathiness can be used to indicate high-priority utterances [50] or dissatisfaction [49]. While the currently available state-of-the-art speech synthesisers do not allow for controlling

Table 4. Scrutinising results

		Conversational prompts	Results Scrutinisation
1.	User:	Hello, I am flying for a conference to Helsinki on the 20th of March. Find me a cheap flight.	*[Results shown on a smartwatch screen]*
2.	CA:	Hello Tom, sure here is a graph displaying flights available on that day, based on price and travel time. The highlighted two seem to be the best tradeoff between price and travel time.	
3.	User:	Are you sure that there are no cheaper options?	
4.	CA:	There em... may be cheaper options, but I don't have access to all service providers. You can check by contacting them.	[Introducing disfluencies and syllable elongation to indicate uncertainty]

breathiness of voice [51], in the future this feature could be used to indicate the involvement of the CA and concern with user's behaviour (e.g. indicating concern about repeated takeaway orders illustrated in Table 2).

On the other hand, disfluencies and elongations (illustrated in Table 4, line 4) can be used to signal CA's uncertainty. Recent research [45] shows that spontaneous (unscripted) audio data can be used to synthesise voices that sound more authentic and better capture the expressive characteristics of speech - this approach offers new possibilities for CA voice design that are relevant to trust calibration.

5 CONCLUSION

In this paper we have bridged insights from the HAI, XAI and IR disciplines to highlight the importance of trust calibration and illustrate the dangers of excessive trust and its implications for user experience. We believe that by providing the user with more control and agency, CA design could foster healthier user-agent relationships. Edwards and Sanoubari [17] emphasised the importance of trust in CAs research and called for its evaluation through the combination of different, inter-disciplinary studies. This call becomes increasingly more important, as the capabilities of CAs are evolving to provide support with more complex tasks that can have financial implications for the users. While we have only proposed four ways of calibrating trust to more appropriate levels, a continued discussion about trust calibration is needed. The CUI community should consider in what meaningful ways trust can be evaluated, in particular focusing on the interplay between ethics and policy making, to foster design of user-centred CAs that promote agency and encourage a long-term use.

ACKNOWLEDGMENTS

This work was supported by the Horizon 2020 FET program of the European Union through the ERA-NET Cofund funding grant CHIST-ERA-20-BCI-001.

REFERENCES

[1] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* (2020).

[2] Kamran Alipour, Jurgen P Schulze, Yi Yao, Avi Ziskind, and Giedrius Burachas. 2020. A study on multimodal and interactive explanations for visual question answering. *arXiv preprint arXiv:2003.00431* (2020).

[3] Jorge A. Alvarado-Valencia and Lope H. Barrero. 2014. Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior* 36 (2014), 102–113. https://doi.org/10.1016/j.chb.2014.03.047

[4] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput. Hum. Interact.* 26, 3 (2019), 17–1.

[5] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. Investigating Proactive Search Support in Conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 1295–1307.

[6] Matthew P Aylett, Selina Jeanne Sutton, and Yolanda Vazquez-Alvarez. 2019. The right kind of unnatural: designing a robot voice. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. 1–2.

[7] Marcia J Bates. 1990. Where should the person stop and the information search interface start? *Information Processing & Management* 26, 5 (1990), 575–591.

[8] Rodrigo Bavaresco, Diórgenes Silveira, Eduardo Reis, Jorge Barbosa, Rodrigo Righi, Cristiano Costa, Rodolfo Antunes, Marcio Gomes, Clauter Gatti, Mariangela Vanzin, Saint Clair Junior, Elton Silva, and Carlos Moreira. 2020. Conversational agents in business: A systematic literature review and future research directions. *Computer Science Review* 36 (2020), 100239. https://doi.org/10.1016/j.cosrev.2020.100239

[9] Jessie Y Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. 2014. *Situation awareness-based agent transparency*. Technical Report. Army research lab aberdeen proving ground md human research and engineering ....

[10] Aleksandr Chuklin, Aliaksei Severyn, Johanne Trippas, Enrique Alfonseca, Hanna Silen, and Damiano Spina. 2018. Prosody modifications for question-answering in voice-only settings. *arXiv preprint arXiv:1806.03957* (2018).

[11] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[12] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. " What can i help you with?" infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.

[13] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331.

[14] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[15] Mateusz Dubiel, Martin Halvey, Leif Azzopardi, Damien Anderson, and Sylvain Daronnat. 2020. Conversational strategies: impact on search performance in a goal-oriented task. In *The Third International Workshop on Conversational Approaches to Information Retrieval*.

[16] Mateusz Dubiel, Martin Halvey, Pilar Oplustil Gallegos, and Simon King. 2020. Persuasive synthetic speech: Voice perception and user behaviour. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–9.

[17] Justin Edwards and Elaheh Sanoubari. 2019. A need for trust in conversational interface research. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. 1–3.

[18] Aaron C Elkins and Douglas C Derrick. 2013. The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents. *Group decision and negotiation* 22, 5 (2013), 897–913.

[19] Xiaocong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater, and Mica R. Endsley. 2008. The influence of agent reliability on trust in human-agent collaboration. *Proceedings of the 15th European conference on Cognitive ergonomics the ergonomics of cool interaction - ECCE '08* (2008), 1. https://doi.org/10.1145/1473018.1473028

[20] Andrew Gibiansky, Sercan Ömer Arik, Gregory Frederick Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep Voice 2: Multi-Speaker Neural Text-to-Speech.. In *NIPS*.

[21] Yunha Han. 2021. Wait, Let's Think about Your Purchase Again: A Study on Interventions for Supporting Self-Controlled Online Purchases. (2021). To Appear in Proceedings of Web Conference 2021.

[22] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570

[23] Philipp Kirschthaler, Martin Porcheron, and Joel E Fischer. 2020. What can i say? effects of discoverability in vuis on task performance and user experience. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–9.

[24] Alexander Kunze, Stephen J. Summerskill, Russell Marshall, and Ashleigh J. Filtness. 2019. Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics* 62, 3 (2019), 345–360. https://doi.org/10.1080/00140139.2018.1547842

[25] J. D. Lee and K. A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[26] Rosemarijn Looije, Mark A Neerincx, and Fokie Cnossen. 2010. Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human-Computer Studies* 68, 6 (2010), 386–397.

[27] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.

[28] D. Manzey, J Elin Bahner, and Anke-Dorothea Hueper. 2006. Misuse of Automated Aids in Process Control: Complacency, Automation Bias and Possible Training Interventions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 3 (2006), 220–224. https:

//doi.org/10.1177/154193120605000303

[29] John M. McGuirl and Nadine B. Sarter. 2006. Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human Factors* 48, 4 (2006), 656–665. https://doi.org/10.1518/001872006779166334 arXiv:https://doi.org/10.1518/001872006779166334 PMID: 17240714.

[30] Bonnie M. Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27, 5-6 (1987), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

[31] M. Ng, K. P. L. Coopamootoo, E. Toreini, M. Aitken, K. Elliot, and A. van Moorsel. 2020. Simulating the Effects of Social Presence on Trust, Privacy Concerns Usage Intentions in Automated Bots for Finance. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*. 190–199. https://doi.org/10.1109/EuroSPW51379.2020.00034

[32] John O'Donovan and Barry Smyth. 2005. Trust in recommender systems. In *Proceedings of the 10th international conference on Intelligent user interfaces*. 167–174.

[33] Thomas C O'Guinn and Ronald J Faber. 1989. Compulsive buying: A phenomenological exploration. *Journal of consumer research* 16, 2 (1989), 147–157.

[34] Richard Pak, Nicole Fink, Margaux Price, Brock Bass, and Lindsay Sturre. 2012. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics* 55, 9 (2012), 1059–1072.

[35] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30, 3 (2000), 286–297.

[36] Claudio S Pinhanez. 2021. Expose Uncertainty, Instill Distrust, Avoid Explanations: Towards Ethical Guidelines for AI. *arXiv preprint arXiv:2112.01281* (2021).

[37] Lingyun Qiu and Izak Benbasat. 2009. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of management information systems* 25, 4 (2009), 145–182.

[38] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction* 37, 1 (2021), 81–96.

[39] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[40] Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. 2016. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 240–249. https://doi.org/10.1109/TVCG.2015.2467591

[41] Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gokhan Tur. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. 41–51.

[42] Ameneh Shamekhi, Q Vera Liao, Dakuo Wang, Rachel KE Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[43] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.

[44] Kimberly Stowers, Nicholas Kasdaglis, Michael Rupp, Jessie Chen, Daniel Barber, and Michael Barnes. 2017. Insights into Human-Agent Teaming: Intelligent Agent Transparency and Uncertainty. In *Advances in Human Factors in Robots and Unmanned Systems*, Pamela Savage-Knepshield and Jessie Chen (Eds.). Springer International Publishing, Cham, 149–160.

[45] Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. 2019. Spontaneous Conversational Speech Synthesis from Found Data.. In *INTERSPEECH*. 4435–4439.

[46] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. 2019. Investigating Users' Preferences and Expectations for Always-Listening Voice Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–23.

[47] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 2018. Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience. In *Proceedings of the Technology, Mind, and Society*. 1–6.

[48] Johanne R Trippas. 2019. *Spoken Conversational Search: Audio-only Interactive Information Retrieval*. Ph.D. Dissertation. PhD thesis, RMIT, Melbourne.

[49] Nigel Ward, Jonathan E Avila, and Aaron M Alarcon. 2021. Towards Continuous Estimation of Dissatisfaction in Spoken Dialog. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 13–20.

[50] Nigel G Ward. 2019. *Prosodic patterns in English conversation*. Cambridge University Press.

[51] Nigel G Ward, Ambika Kirkland, Marcin Włodarczak, and Eva Székely. [n.d.]. Two Pragmatic Functions of Breathy Voice in American English Conversation. ([n. d.]).

[52] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2020. "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces* (2020), 1–12.

[53] Ryen W White and Ian Ruthven. 2006. A study of interface support mechanisms for interactive information retrieval. *Journal of the American Society for Information Science and Technology* 57, 7 (2006), 933–948.

[54] Christine T. Wolf and Kathryn E. Ringland. 2020. Designing Accessible, Explainable AI (XAI) Experiences. *SIGACCESS Access. Comput.* 125, Article 6 (March 2020), 1 pages. https://doi.org/10.1145/3386296.3386302

[55] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256* (2016).

[56] Sunghwan Yi and Hans Baumgartner. 2011. Coping with guilt and shame in the impulse buying context. *Journal of Economic Psychology* 32, 3 (2011), 458–467.

[57] Debora Zanatto, Massimiliano Patacchiola, Jeremy Goslin, and Angelo Cangelosi. 2016. Priming anthropomorphism: Can the credibility of humanlike robots be transferred to non-humanlike robots?. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 543–544.