Automating Crisis Communication in Public Institutions – Towards Ethical Conversational Agents That Support Trust Management

Lennart Hofeditz
University of Duisburg-Essen, Germany, lennart.hofeditz@uni-due.de

Milad Mirbabaie
Paderborn University, Germany, milad.mirbabaie@uni-paderborn.de

Lukas Erle
University of Duisburg-Essen, Germany, lukas.erle@uni-due.de

Eileen Knoßalla
University of Duisburg-Essen, Germany, eileen.knossalla@uni-due.de

Lara Timm
University of Duisburg-Essen, Germany, lara.timm@uni-due.de

Automating Crisis Communication in Public Institutions – Towards Ethical Conversational Agents That Support Trust Management

Lennart Hofeditz[1], Milad Mirbabaie[2], Lukas Erle[1], Eileen Knoßalla[1], Lara Timm[1]

[1] University of Duisburg-Essen, Department of Computer Science and Applied Cognitive Science, Duisburg, Germany
{lennart.hofeditz, lukas.erle, eileen.knossalla, lara.timm}@uni-due.de
[2] Paderborn University, Department of Information Systems, Paderborn, Germany
{milad.mirbabaie}@uni-paderborn.de

Abstract. To improve disaster relief and crisis communication, public institutions (PIs) such as administrations rely on automation and technology. As one example, the use of conversational agents (CAs) has increased. To ensure that information and advisories are taken up seriously, it is important for PIs to be perceived as a trusted source and a trustworthy point of contact. In this study, we therefore examine how CAs can be applied by PIs to, on the one hand, automate their crisis communication and, on the other hand, maintain or even increase their perceived trustworthiness. We developed two CAs – one equipped with ethical cues in order to be perceived more trustworthy and one without such cues – and started to conduct an online experiment to evaluate the effects. Our first results indicate that applying ethical principles such as fairness, transparency, security and accountability have a positive effect on the perceived trustworthiness of the CA.

Keywords: Public institutions, Conversational agents, Corona crisis, Trust, AI.

1 Motivation

Crises, such as the Covid-19 pandemic or the flood disaster in West Germany in 2021, pose major challenges for public institutions (PIs) such as municipal administrations or public facilities [1]. They are required to both react quickly to recent developments and publish accurate information. In such challenging and unforeseen situations, efficient crisis communication is necessary to mitigate damage and protect human lives [2–4]. PIs play an important role in these crisis situations as they usually coordinate and manage the communication regarding the crisis [5]. However, PIs often lack resources which slows down their response time and results in people's sudden information needs being not met [6]. To improve disaster relief and the response time in their crisis communication, the use of automation and communication technologies is increasingly being considered [2, 7, 8]. One group of technologies that is increasingly being applied in this context are conversational agents (CAs). CAs include systems that provide an

17th International Conference on Wirtschaftsinformatik,
February 2022, Nürnberg, Germany

enjoyable user experience [9] by interacting with people in natural language via text or voice [10]. CAs can include self-learning capabilities via artificial intelligence (AI)-based machine learning algorithms [11]. One example in the context of Covid-19 was the CA "COVINFO" [8] that was tested to provide current information on the pandemic. However, it is often not visible to the user where the information provided by the CA originates from and who is responsible for the system, which in turn can result in a decrease of trust in the PIs, which have a certain obligation to follow ethical principles such as transparency, accountability, explainability, fairness and security [12]. We therefore address this issue by examining how the perceived trustworthiness of PIs can be improved by using CAs that follow certain ethical principles in crisis communication. This led us to the following research question:
RQ: *How can ethical principles be applied to conversational agents in order to increase trustworthiness of public institutions during crisis events*?

We developed two CAs – one containing the ethical principles *fairness*, *transparency*, *accountability*, *security & data privacy* and *autonomy* derived from relevant literature and one without these cues – and started to conduct an online experiment. We implemented the study as a 2x1 between-subjects-design, in which the participants have either interacted with an "ethical" or a "neutral" CA that provided information about the status of the Covid-19 pandemic. We derived hypotheses to test how different ethical cues are connected to trustworthiness and to provide knowledge on how ethical principles can increase the trust in CAs and in PIs during a crisis.

2 Background and Hypotheses

CAs can be defined as any dialogue system that uses natural language processing and automatically responds in human language [9, 13, 14]**.** In addition, they can use machine intelligence (MI) in order to respond to all possible actions of the user. MI "was defined as the ability of a trained computer system to provide rational, unbiased guidance in such a way that achieves optimal outcomes in a range of environments and circumstances" [15]. In our study, CAs are implemented as text-based systems which are used as stimulus material for an online experiment. For this, we used DialogFlow from Google[1]. It can therefore be described as a conversational AI that is able to understand and learn from received messages through training phrases [16]. As a leader in conversational AI algorithms, Google addresses machine learning fairness in their products in order to prevent biases. These training phrases have been prepared by training the CA with certain "trigger words" that prompt different responses. Furthermore, our CA uses natural language processing (NLP) through sensitive input processing, pre-defined potential meanings and different responses [17]. CAs can be equipped with different social cues to be perceived more human-like [18, 19]. Although CAs have the technical abilities to support the crisis communication of PIs, they also need to be perceived as a trustworthy tool and a trusted source of information to ensure that they are used and helpful. Trust is a well-known concept and has been strongly

[1] https://cloud.google.com/dialogflow/es/docs/training

discussed in past research. It is described as the "willingness to be vulnerable to another party based on the belief that the latter party is 1) competent, 2) open, 3) concerned, and 4) reliable" [20]. A differentiation is made between trust in people and trust in technology [21]. Furthermore, trustworthiness represents one of the three main goals when trying to achieve human-centered AI (HCAI) [22]. In this context, the European Independent High-Level Expert Group on AI classified three components of a trustworthy AI: Applicable laws and regulations, compliance of ethical principles and technical robustness [23]. Trustworthiness is seen as an overarching ethical principle [16, 22]. In the present study, we focus on trust in PIs and to this end developed a prototype of a trustworthy CA that can provide crisis related information. In accordance with the social cues for CAs developed by Feine et al. [19] and considering a context of European PIs, we equipped the "ethical" CA with certain ethical cues regarding the elaboration of the ethical principles of trustworthy AI [23] Considering current research, the ethical principles of fairness, transparency, security & data privacy, accountability, and respect for human autonomy were attributed a very high relevance especially in the European context [16, 23, 24]. In order to examine how these ethical cues can increase the trust in a CA, we derived the following hypotheses: [H1-5] *The perceived {fairness, transparency, accountability, security & data privacy, respect for human autonomy} of a CA has a positive influence on the perceived trustworthiness*.

3 Study Design and Preliminary Results

To answer the research question and test our hypotheses, we developed two CAs that differ in the fact that one of them (named "German Health Assistant") offers a variety of ethical cues and social cues such as those proposed by Feine et al. [19] with regard to ethical principles from previous research [16, 23, 24], while the second CA (named "Covid Assistant") is not equipped with these cues and therefore labeled as the "neutral" CA. However, both CAs use the same source for information and follow a largely similar conversational pattern. The only difference between the two CAs were the present/absent ethical cues. The source for all information is offered by the German national ministry of health on the website *zusammengegencorona.de* [25]. Ethical cues of the German Health Assistant include but are not limited to the features in Table 1.
To improve the users' understanding of the agent's competences, we added a series of Covid-19 related topics (for example "basic knowledge about the coronavirus") to choose from at the beginning of the chat. This guidance is supposed to narrow down the possible triggers and their respective responses. The CAs have furthermore been integrated into two websites that were designed slightly different: The website for the ethical CA has been styled to mimic an official website of the German Robert-Koch-Institute. The agent itself has also been styled accordingly.

Table 1. Implementation of the social cues for the German Health Assistant

Ethical principles	**Ethical cues**
Fairness	using gender-neutral language wherever possible

Security & data privacy	GDPR-conformity disclaimers
Transparency and accountability	Providing links with additional information about the content and the CA
Respect for human autonomy	Asking to start the dialogue, and waiting for user input

The website for the neutral CA has been designed to look more informal and unofficial. To test the effect of the ethical cues of the German Health Agent CA on the perceived trustworthiness, we conducted a quantitative 2x1 between-subjects-design online study consisting of an interaction task with a CA platform (for example gathering information on the virus mutations) and some online questionnaires focusing on the perception of the ethical cues, the PI and trustworthiness of the CA. To determine whether there were any technical difficulties with the interface or interaction with the CA, we conducted a pretest with N = 10 participants. We also used the pretest to validate self-developed question scales on the ethical cues and excluded two items to improve reliability. Furthermore, we used the pretest to briefly evaluate the sufficiency of the ethical cues. Overall, *fairness* was measured with six items. One example item was " The chatbot is free of bias". *Transparency* was measured by eight items. An example item was: "The chatbot makes it clear where it retrieves its information from." *Security and data privacy* was measured with nine items. An example item was: "The chatbot prevents unauthorized access to data." Accountability was measured with four items. An example item was "The chatbot provides the ability to report problems with the chatbot." *Respect for human autonomy* was measured with seven items. An example item was "The chatbot does not take decision-making away from users."

Participants for the main study have been recruited from different online communities via email and social media. They randomly interacted with either the ethical or the neutral CA. Irrespective of which CA the participants have seen, both groups received the same set of three tasks, which all were spread across different topics within the pandemic and all required interaction with the agent. After completing the interaction, participants were asked to name the organization that offered the CA and whether they think they can trust this organization. As a next step, participants were asked to judge the perceived trustworthiness of the CA they interacted with. To evaluate the CA in regard to the ethical principles, the items that had been validated in the pretest and developed from the findings of [18, 23, 24] were inquired using a 7-point Likert scale. Afterwards**,** the human-computer trust scale [26] was used to measure the perceived trustworthiness of the CA in more detail. We further added the item “I can trust the chatbot” to this scale, which was also measured using a 7-point Likert scale. For additional evaluation of the CAs, the perceived usefulness and intention to use were measured and finally, participants were asked to enter their demographics.

We have already gathered preliminary data from 157 participants and started to analyze 101 datasets. To measure the influence of the individual ethical principles on perceived trustworthiness, we conducted a first linear regression analysis with the independent variable being the ethical principles and the dependent variable being overall trustworthiness). The model proved to be significant (F (5,95) = 42.96, p < .001)

with a reasonably high regression accuracy of R = .83 and R^2 = .68. The standard mean error of this analysis had a value of *SE* = .58.

Table 2. Results of the linear regression analysis

Variable	**B**	***T* (99)**	***p***
(Constant)	-2.31	-6.28	.000
Fairness	.15	2.06	.000
Transparency	.26	2.63	.042
Security & Data Privacy	.48	4.41	.010
Accountability	.20	2.56	.012
Respect for Human Autonomy	-.02	-.32	.748

As shown in Table 2, all ethical principles except *respect for human autonomy* showed a significance of $p < .05$. This indicates that the hypotheses H1 to H4 can be supported while hypothesis H5 should be rejected.

4 First Conclusions and Next Steps

Our preliminary results on the effect of the ethical cues on trustworthiness support previous literature [18, 23, 24]. However, the ethical cue *respect for human autonomy* seems to have no significant effect on the trust in the CA and the PI. We therefore conclude that a CA that uses a more dominant language is more likely to be trusted [27]. This might be explained by the fact that in crisis situations, citizens rely on clear guidance from PIs. Furthermore, a catalogue of ethical cues especially focusing on *fairness, transparency, security and accountability* and their practical implementations could offer best practices for developing trustworthy CAs. Regarding next steps, we will collect more data to achieve a broader sample size. We will further re-evaluate the existing data and test our hypotheses, as well as evaluate the influence of the perceived trustworthiness on the perceived usefulness and intention to use such a system.

We expect that the use of ethical cues will successfully increase the ethical principles conveyed by the agent, which in turn will increase the perceived trustworthiness of the CA. Furthermore, we expect to be able to validate the ethical cues we derived from existing literature and aim to identify and analyze more ethical cues in the context of PIs. Beyond that, an analysis regarding which principles generate more or less trust and whether the design of ethical cues themselves has any meaningful impact could be beneficial to both PIs and the scientific community. For this, it is also important to examine the ethical cues' effect without the presented websites in order to exclude the effect of biases.

Our research offers guidance for PIs on how they can use CAs to communicate information regarding crises both quickly and accurately, as well as in line with ethical standards and norms to maintain or possibly increase their trustworthiness. This research will help extend the knowledge on the perception of ethical CAs as well as how ethical principles are intertwined and how they can be tested.

References

1. Quinn, P.: Crisis Communication in Public Health Emergencies: The Limits of 'Legal Control' and the Risks for Harmful Outcomes in a Digital Age. Life Sci. Soc. Policy. 14, 1–31 (2018). https://doi.org/10.1186/s40504-018-0067-0.
2. Hofeditz, L., Ehnis, C., Bunker, D., Brachten, F., Stieglitz, S.: Meaningful Use Of Social Bots? Possible Applications In Crisis Communication During Disasters. In: European Conference on Information Systems. AIS Electronic Library, Stockholm (2019).
3. Stieglitz, S., Bunker, D., Mirbabaie, M., Ehnis, C.: Sense-making in social media during extreme events. J. Contingencies Cris. Manag. 26, 4–15 (2018). https://doi.org/10.1111/1468-5973.12193.
4. Mirbabaie, M., Bunker, D., Stieglitz, S., Marx, J., Ehnis, C.: Social media in times of crisis: Learning from Hurricane Harvey for the coronavirus disease 2019 pandemic response. J. Inf. Technol. 35, 195–213 (2020). https://doi.org/10.1177/0268396220929258.
5. Boin, A., Lodge, M.: Designing resilient institutions for transboundary crisis management: A time for public administration. public Adm. 94, 289–298 (2016).
6. Wang, J., Hutchins, H.M.: Crisis Management in Higher Education: What Have We Learned From Virginia Tech? Adv. Dev. Hum. Resour. 12, 552–572 (2010). https://doi.org/10.1177/1523422310394433.
7. Fan, C., Zhang, C., Yahja, A., Mostafavi, A.: Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management. Int. J. Inf. Manage. 56, 102049 (2021). https://doi.org/10.1016/j.ijinfomgt.2019.102049.
8. Maniou, T.A., Veglis, A.: Employing a chatbot for news dissemination during crisis: Design, implementation and evaluation. Futur. Internet. 12, 1–14 (2020). https://doi.org/10.3390/FI12070109.
9. Diederich, S., Brendel, A.B., Kolbe, L.M.: On Conversational Agents in Information Systems Research: Analyzing the Past to Guide Future Work. Proc. Int. Conf. Wirtschaftsinformatik. (2019).
10. McTear, M., Callejas, Z., Griol, D.: The Conversational Interface. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-32967-3.
11. Bittner, E., Oeste-Reiß, S., Leimeister, J.M.: Where is the Bot in our Team? Toward a Taxonomy of Design Option Combinations for Conversational Agents in Collaborative Work. In: Proceedings of the 52nd Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences (2019). https://doi.org/10.24251/hicss.2019.035.
12. EU: Ethics guidelines for trustworthy AI, https://ec.europa.eu/futurium/en/ai-alliance-consultation, last accessed 2021/07/29.
13. Tavanapour, N., Poser, M., Bittner, E.A.C.: Supporting the idea generation process in citizen participation - Toward an interactive system with a conversational agent as facilitator. 27th Eur. Conf. Inf. Syst. - Inf. Syst. a Shar.

Soc. ECIS 2019. 0–17 (2020).
14. Pfeuffer, N., Benlian, A., Gimpel, H., Hinz, O.: Anthropomorphic Information Systems. Bus. Inf. Syst. Eng. 61, 523–533 (2019). https://doi.org/10.1007/s12599-019-00599-y.
15. Cutillo, C.M., Sharma, K.R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K.D.: Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit. Med. 3, (2020). https://doi.org/10.1038/s41746-020-0254-2.
16. Floridi, L.: Establishing the rules for building trustworthy AI. Nat. Mach. Intell. 1, 261–262 (2019). https://doi.org/10.1038/s42256-019-0055-y.
17. Rybakova, M.: How to Build Smarter Bots With AI: Agencies, https://chatfuel.com/blog/posts/build-ai-chatbots, (2020).
18. Wambsganß, T., Höch, A., Zierau, N., Söllner, M.: Ethical Design of Conversational Agents: Towards Principles for a Value-Sensitive Design. 16th Int. Conf. Wirtschaftsinformatik. (2021).
19. Feine, J., Gnewuch, U., Morana, S., Maedche, A.: A taxonomy of social cues for conversational agents. Int. J. Hum. Comput. Stud. 132, 138–161 (2019). https://doi.org/10.1016/j.ijhcs.2019.07.009 Please.
20. Mishra, A.K.: Organizational Responses to Crisis: The Centrality of Trust. In: Trust in Organizations: Frontiers of Theory and Research. pp. 261–287. SAGE Publications, Inc, Thousand Oaks: California (1996). https://doi.org/10.4135/9781452243610.n13.
21. Thiebes, S., Lins, S., Sunyaev, A.: Trustworthy artificial intelligence. Electron. Mark. 31, 447–464 (2020). https://doi.org/10.1007/s12525-020-00441-4.
22. Shneiderman, B.: Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans. Interact. Intell. Syst. 10, (2020). https://doi.org/10.1145/3419764.
23. Independent High-Level expert Group on Artifical Intelligence: Ethics Guidelines for Trustworthy AI: Set up by the European Commission. (2019).
24. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., Vayena, E.: AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach. 28, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5.
25. Bundesministerium für Gesundheit: Zusammen gegen Corona, https://www.zusammengegencorona.de/, (2021).
26. Gulati, S.N., Sousa, S.C., Lamas, D.: Design, development and evaluation of a human-computer trust scale. Behav. Inf. Technol. 38, 1004–1015 (2019). https://doi.org/10.1080/0144929X.2019.1656779.
27. Zhou, M.X., Mark, G., Li, J., Yang, H.: Trusting Virtual Agents. ACM Trans. Interact. Intell. Syst. 9, 1–36 (2019). https://doi.org/10.1145/3232077.