Computers and Society Research Journal (2021), **1**
https://doi.org/10.54822/QXOM4114

ARTICLE

Applying Ethical AI Frameworks in practice: Evaluating conversational AI chatbot solutions

Suzanne Atkins, Ishwarradj Badrie, and Sieuwert van Otterloo

Email: sieuwert@ictinstitute.nl

(First published online September 24, 2021)

Abstract

Ethical AI frameworks are designed to encourage the accountability, responsibility and transparency of AI applications. They provide principles for ethical design. To be truly transparent, it should be clear to the user of the AI application that the designers followed responsible AI principles. In order to test how easy it is for a user to assess the responsibility of an AI system and to understand the differences between ethical AI frameworks, we evaluated four commercial chatbots against four responsible AI frameworks. We found that the ethical frameworks produced quite different assessment scores. Many ethical AI frameworks contain requirements/principles that are difficult to evaluate for anyone except the chatbot developer. Our results also show that domain-specific ethical AI guidelines are easier to use and yield more practical insights than domain-independent frameworks. We conclude that ethical AI researchers should focus on studying specific domains and not AI as a whole, and that ethical AI guidelines should focus more on creating measurable standards and less on stating high level principles.

Keywords: Artificial Intelligence; Responsible AI; Ethics; Guidelines; Chatbots; Compliance; Conversational AI

Artificial intelligence (AI) applications grow more complex every year. However, many aspects of AI raise potential ethical issues, including the privacy and security of personal data (1; 2), model bias and fairness (3; 4), and the consequences of using machines to make decisions that affect individuals and society (5).

The development of responsible AI is a research agenda that focuses on the development and implementation of ethical, transparent and responsible solutions to mitigate trust and privacy issues (6). Responsible AI focuses on the human-user aspects of the system and requires compliance with stakeholder expectations, laws and regulations, and incorporates social and ethical standards. The fundamental principles of responsible AI are accountability, responsibility and transparency (ART) (7):

- **Accountability** requires that the system can explain and defend its decisions and actions. To be accountable, the results of decision making algorithms must be communicable. The system design must also be accountable for upholding the moral ideals and social norms inherent in the system working environment.
- **Responsibility** of individuals and the ability of AI systems to react and detect errors or unforeseen consequences in response to decision-making.
- **Transparency** or explainability of the system requires that methods by which AI systems make judgments and learn to adapt to their environment and regulate the information used or generated must be described, inspected and replicated.

Many frameworks have been designed to guide the development of responsible AI. However, the grand philosophical principles of responsible AI (7; 8) from some of these frameworks are not readily applicable when designing small AI tools. Domain specific toolkits and frameworks have therefore been developed for applications such as gaming (9) and conversational AI (10). These include many of the principles of responsible AI, but in a more directly applicable manner, focusing on good design.

One of the key tenets shared by all these frameworks is transparency. In order to be fully transparent, the responsible design of the AI must be clearly visible to the user. In this study, we study chatbots from an outside perspective (without access to the source code or internal documentation), to explore to what extent decision makers can assess the quality and ethics of a deployed AI using four different responsible AI frameworks.

Chatbots

Chatbots are a common part of customer service interactions, appearing on websites from supermarkets to government organisations. An screenshot from a British bank can be seen in figure 1. This chatbot was not used in the study.

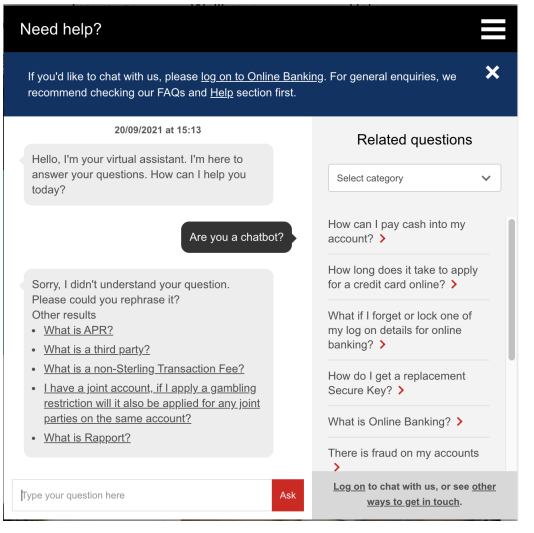

Figure 1. An example of customer service chatbot from a British bank. This was not one of the chatbots used in this study.

The modern concept of a chatbot began with Alan Turing's Imitation Game or the Turing test (11), where a computer aims to mimic human behaviour. ELIZA, the first usable chatbot, was created in 1966 (12). This system employed keyword matching and basic context identification, but ELIZA was a basic system that was incapable of maintaining a dialogue between people and bots. The development of natural language processing (13) and the creation of the ALICE (Artificial Linguistic Internet Computer Entity) chatbot (14) marked the birth of the chatbots we interact with today (15). Chatbots are often trained using human service conversations and chatbots are often used in conjunction to human-to-human chat.

Chatbots are one of the most visible ways that consumers interact with AI tools. In the best situation, users may view them as a good way of received immediate support, even outside officer hours. They may also v frustrating obstacle to customer support, they are also ambassadors for AI in daily life, and particularly in customer service. They are therefore an excellent subject to consider how ethical considerations and responsible AI should be applied to applications that users encounter on a daily basis.

Chatbots raise several potential ethical issues. The chatbot is trained on data, generally drawn from past user interactions, and may continue to gather and learn from data provided during deployment. This introduces potential issues about bias in the training data that could lead to incorrect information, biased responses or insulting language. It also raises security and privacy issues, depending on what and how data are stored, and how or if permission is sought. At a more abstract level, data introduces a potential imbalance in information between user and bot. The chatbot potentially has access to information ranging from the user's address, location or the device used to an extensive database about the user. This could be considered to put the user at a disadvantage if there is an information imbalance. Designers of responsible chatbots must also contend with

questions about how to minimise inbuilt bias in the models, and what processes are in place to identify and prevent incorrect behaviour, whether that be discrimination, offensive language, or simply giving the wrong answers or advice.

Frameworks for Responsible AI

In this work, we assess publicly deployed chatbots against four modern frameworks for responsible AI: the DEON Checklist (16); the Artificial Intelligence Impact Assessment (AIIA) (17); the Assessment List for Trustworthy Artificial Intelligence (ALTAI) (18); and Microsoft's conversational AI guidelines (10). The first three sets of guidelines are designed to apply to general AI applications. The Microsoft guidelines are the only ones specifically designed for conversational AI, of which chatbots are a subclass. The principles within the frameworks can be categorised according to type, as shown in figure 2.

Figure 2. Classification of types of guideline within each framework. Transparency includes auditability and statement of intent; culture includes environmental and social factors; reliability includes testing metrics; fairness includes bias in data sets and participation of stakeholders; privacy and security include data gathering and storage considerations; accessibility is how easy the AI is to use.

The DEON checklist for the responsible use of data in data science projects is part of a command-line application for adding ethical checklists to data science projects (16) available from deon.drivendata.org. The checklist from DEON in this study includes 21 checkpoints grouped into five phases: data collection, data storage, analysis, modelling, and deployment. The checklist has been developed by considering real life incidents that could have been avoided with appropriate planning and ethical consideration.

The AIIA assists in identifying legal and ethical questions arising from the use of AI, taking into account the appropriate framework of standards and trade-offs (17) available from ecp.nl. It builds on the 2006 "Guidance on Conduct for Autonomous Systems" (8), which focused on the legal aspects of implementing autonomous systems. The AIIA was written by experts from a variety of fields, including lawyers, management experts and technicians and is published by the Dutch Platform for the Information Society (ECP). The AIIA is broader than the DEON checklist and includes ethical aspects. The working group considers that AI should improve well-being and not only respect but also promote human values, as is clear in figure 2. However, not all of the principles are directly relevant to chatbots.

The ALTAI was established by the High-Level Expert Group on AI, which is part of the European Commission's digital strategy (18) available from the digital strategy library (link). The ethics rules for trustworthy AI are the most comprehensive of the four considered. They require that an assessment list be used in order to assess whether or not the AI system being considered corresponds to seven standards of trustworthy artificial intelligence including 69 sub guidelines. Like

the AIIA, these include broad principles about the contribution of AI to societal development and well-being, as well as adherence to legislation and standards.

Microsoft's conversational AI guidelines are the only ones considered in this study that were specifically designed for systems such as chatbots. They are intended to assist in designing a bot that fosters trust in the organisation and service (10), and are available from www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_Nov_2018.pdf. They consist of 28 guidelines, published in November 2018 by the Microsoft Corporation, with the intention of focusing on transparency among organisations in the use of AI (19).

Testing the Frameworks on Chatbots

We selected chatbots from four different Dutch organisations. Each was from a different sector: banking, healthcare insurance, telecommunications, and government. As users, we attempted to test the principles from each framework, by checking the supporting privacy and security statements and using the chatbots directly. For each principle, we can answer yes or no/unknown. Unknown meant that we could not determine as external decision makers whether this principle was met. In this way, we test how clear it is to decision makers outside the development team that the chatbot was developed in line with responsible AI principles. We also test the effectiveness of each guideline framework.

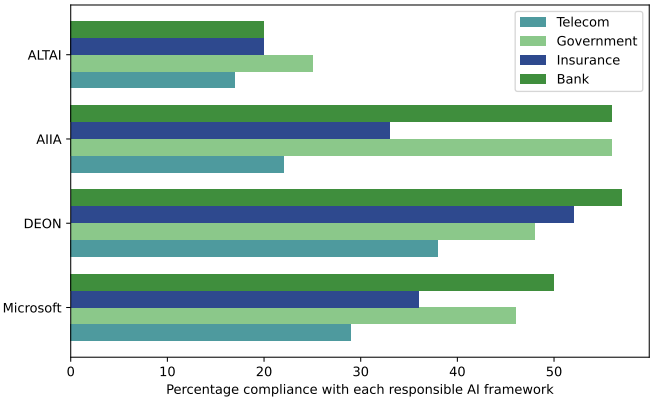

Figure 3. Apparent compliance with four responsible AI frameworks, as assessed by user.

In figure 3, we show the percentage of the principles from each set of guidelines that we can ascertain to have been met as users of the chatbot. The chatbots, which are all already deployed, get low scores from the ALTAI framework. This is because a large number of the guidelines pertain to features that are untestable by outside decision makers, such as resilience to attack, auditability, risk management and impact on environmental well-being. Some of the factors included in the ALTAI guidelines, such as general safety and impact on democracy were written with more powerful AI systems in mind. We expect some of these factors to have been considered by the developers under data privacy laws and security management, however as users we see no evidence of this. We also find that many of these criteria are more philosophical and are therefore not helpful if a user wants to assess the responsible design of a chatbot.

The chatbots from the bank and government both perform better on the Microsoft and AIIA frameworks. This is because both of these chatbots were accompanied by very clear privacy statements setting out data use and storage. Again, we expect the other companies have data manage-

ment plans under European privacy (GDPR) legislation, however they did not communicate this clearly to the user. These bots also made it clearer that they were in fact bots and that their ability to answer questions was limited.

We could not thoroughly test the bias of the chatbots and therefore none of them scored highly against the bias guidelines. All of them were programmed to shut down when presented with racist or abusive language, however we did not test if they performed differently when more subtle profiling markers were used, for example if language indicative of an immigrant background was used. The chatbots are not able to take business decisions, instead handing over to human operators. The risk from bias is therefore low and is limited to poor performance when answering questions.

Conclusion

The current generation of responsible AI guidelines are a poor way to assess the ethics of chatbots as a user. Many of the principles are vague, hard to understand as an average user and often irrelevant to chatbots. Chatbot specific frameworks, such as that from Microsoft may be a useful tool for developers, and are easier to apply as a user than more general philosophical frameworks. However, much more transparency is required so that an average user can see that the chatbot is compliant. Due to lack of transparency, we found that frameworks we tested were not useful tools for users to test the ethical design of chatbots.

Acknowledgement

This article includes results from the thesis 'Measuring the Responsible Use of Conversational AI: The Application of AI Frameworks on Chatbots' by I. Badrie, submitted for his BSc in Information Sciences at the Vrije Universiteit Amsterdam, The Netherlands.

The scores and a summary table of each framework are available at https://ictinstitute.nl/computers-society-research-journal/. The logo is save earth by Laymik from the Noun Project.

References

[1] Chin-Lung Hsu and Judy Chuan-Chuan Lin. An empirical examination of consumer adoption of Internet of Things services: Network externalities and concern for information privacy perspectives. *Computers in Human Behavior*, 62:516–527, 2016, https://doi.org/10.1016/j.chb.2016.04.023.

[2] Bernd Carsten Stahl and David Wright. Ethics and privacy in ai and big data: Implementing responsible research and innovation. *IEEE Security & Privacy*, 16(3):26–33, 2018, https://doi.org/10.1109/MSP.2018.2701164.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

[4] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021, https://doi.org/10.1145/3457607.

[5] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017, https://doi.org/10.1609/aimag.v38i3.2741.

[6] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. Principles to practices for responsible AI: Closing the gap. *arXiv preprint*, 2020, arXiv:2006.04707.

[7] Virginia Dignum. Responsible artificial intelligence: designing AI for human values. *Daffodil International University*, 2017.

[8] ECP Platform voor de Informatiesamenleving. Handreiking voor gedragsregels Autonome Systemen. Juridische aandachtspunten voor de bouw en het gebruik van autonome systemen. leidschendam: ECP.nl, 2006. https://ecp.nl/wp-content/uploads/2017/04/Handreiking-voor-gedragsregels-autonomous-systems-2006.pdf.

[9] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2018.

[10] Microsoft Research. Responsible bots: 10 guidelines for developers of conversational AI., 2018. https://www.microsoft.com/en-us/research/publication/responsible-bots/, accessed on 2021/09/15.

[11] Alan M Turing. Computing machinery and intelligence. In Robert Epstein, Gary Roberts, and Grace Beber, editors, *Parsing the Turing test*, pages 23–65. Springer, 2009.

[12] Joseph Weizenbaum. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966, https://doi.org/10.1145/365153.365168.

[13] Gobinda G Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003, https://doi.org/10.1002/aris.1440370103.

[14] Wallace R.S. The Anatomy of A.L.I.C.E. In Robert Epstein, Gary Roberts, and Grace Beber, editors, *Parsing the Turing test*. Springer, 2009.

[15] Carlene Lebeuf, Alexey Zagalsky, Matthieu Foucault, and Margaret-Anne Storey. Defining and classifying software bots: A faceted taxonomy. In *2019 IEEE/ACM 1st International Workshop on Bots in Software Engineering (BotSE)*, pages 1–6. IEEE, 2019.

[16] DEON. An ethics checklist for data scientists, 2018. https://deon.drivendata.org/#default-checklist, accessed on 2021/09/15.

[17] ECP Platform voor de Informatiesamenleving. Artificial Intelligence Impact Assessment (English version), 2020. https://ecp.nl/publicatie/artificial-intelligence-impact-assessment-english-version/, accessed on 2021/09/15.

[18] European Commission: High-Level Expert Group on Artificial Intelligence. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment: Shaping Europe's digital future., 2020. https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment, accessed on 2021/09/15.

[19] Lili Cheng. The Official Microsoft Blog: Microsoft introduces guidelines for developing responsible conversational AI, 2018. https://blogs.microsoft.com/blog/2018/11/14/microsoft-introduces-guidelines-for-developing-responsible-conversational-ai/, accessed on 2021/09/15.