Ethics Check for Conversational AIs

Sophie Hundertmark
Lucerne University of Applied
Sciences and Arts
6002 Lucerne
Switzerland
sophie@hundertmark.ch

Edy Portmann
University of Fribourg
1700 Fribourg
Switzerland
edy.portmann@unifr.ch

Abstract

We are developing an ethics check for chat and voicebots. Inspired by the Turing test, which is supposed to find out how human-correct an AI is, we want to identify ethically correct bots and give hints for optimization to those that are not yet. To define ethically correct bots as such, we first must define what ethically correct means in our language and culture. Subsequently, we must define measurement values, how the ethical criteria developed before can be measured. This is followed by an assessment, which each bot can undergo to obtain an evaluation of the state of its ethical correctness, including any potential for improvement. The assessment will initially be conducted by humans, but it is conceivable that bots will also be able to solve the assessment in the future.

Author Keywords

Bots, Chatbots, Conversational Agents, Conversational AI, Conversation Theory, Computational Ethics, Digital Ethics, Fuzzy Ethicity, Life Engineering

CSS Concepts

• **Human-centered computing; Human computer interaction (HCI);** HCI design and evaluation methods; Usability testing

Introduction

Recently, influenced by the technological and scientific advances in the fields of artificial intelligence (AI) and machine learning (ML) and by the growing acceptance of non-human communication partners, the number of companies using chatbots or conversational agents (CAs) to automate their customer touchpoints has grown [1]. In general, AI-based CAs, such as Amazon's Alexa or Apple's Siri, have become a dominant service interface between providers and users. These CAs are designed with the intention of supporting users in their everyday lives as intelligent personal assistants [2]. Chatbots simulate human communication and can better take on human characteristics compared to other software-based programs [3]. Therefore, their high degree of interaction capability is expected to affect the user experience and behaviour. Hence, to improve this type of communication technology, companies invest a large amount of effort in the technical development of

CAs [1]. To test whether a conversational agent is as human as a discussion partner companies may apply the already existing Turing test. This test was developed in 1950 by Alan Turing and describes a way to test the intelligence of machines. In the Turing test, a human questioner converses with two to three other interlocutors [4]. The conversation only takes place via chat, so the interface is a classic computer with screen and keyboard. The interlocutors cannot see each other. The interesting thing about the conversation is that one of the interlocutors is a machine and the other one or two are real people. The questioner does not know behind which interlocutor the machine is. The human questioner has the task to ask intensive questions. At the end he must decide which of his interlocutors is a human or a machine. If the questioner does not clearly find out who is a machine, the chatbot or the machine has passed the Turing test. At the same time, discussions about ethical standards for all kinds of technology become more popular. Companies invest a lot of time in a balanced understanding of digital ethics and sustainability concepts [5]. As a result, more and more organizations have started to implement ethical guidelines within their own company structure. According to Oesterle ethics is essential in the context of life engineering and only if we integrate ethics into our technological developments, we will achieve the highest quality of life for everyone [6]. Floridi argues that efforts in this regard should also go beyond hard ethics and consider so-called soft ethics. Soft ethics is about thinking, in addition to fixed regulations and technical restrictions, about what should and should not be done beyond the existing norms and what everyone can contribute in terms of self-regulation [7].

To our knowledge and research efforts, that we made, however, there is no known test that also checks conversational AIs for ethical standards. We want to close this gap with our work. Therefore, we ask the following research questions:

RQ1: Which are the ethical guidelines that a conversational AI should have?

RQ2: How can we measure these ethical guidelines (its ethicity [1]) of a conversational AI?

RQ3: How can an AI use the pre-defined ethical guidelines to conduct an ethical check autonomously without any human support?

Our General Idea

We are developing an ethics test for chat- and voicebots resp. conversational AIs. Inspired by the Turing test, which is supposed to find out how human-correct an AI is, we want to identify ethically correct bots and give hints for optimization to those that are not yet. To define ethically correct bots as such, we first must define what ethically correct means in our language and culture. And we must consider that these values might change over time. Subsequently, we must define measurement values, how the ethical criteria developed before can be measured. Here we considering Scheler's research “Wertethik”. According to Scheler (1973), "all that is good" is based on the realisation of the good and on turning in on oneself. For

[1] Ethicity means a fuzzy measuring of an ethical value.

Scheler (1973), the value phenomena that can be experienced intuitively have a similar character and function. After the measurements are defined we develop an assessment, which each bot can undergo to obtain an evaluation of the state of its ethical correctness, its ethicity, including any potential for improvement. The assessment will initially be conducted by humans, but it is conceivable that bots will also be able to solve the assessment in the future.

The ethics check for conversational agents is doubly novel. On the one hand, there are no recognized or widely used ethical guidelines for conversational AI projects, neither in research nor in practice. On the other hand, apart from the Turing test, there are no widely used benchmarks or other tests that evaluate conversational agents while also identifying optimization opportunities. Our project even combines both aspects in a final application. In contrast to the Turing test, in which a human chat with a conversational AI, our ethics check should be able to be performed directly by a chatbot at the end. The chatbot then has the defined set of ethics rules, knows which questions to ask, and can match and rank the answers of the chatbot under test with the benchmarks we have defined. The result is a fully automated ethics check that is also transparent, as it shows exactly which criteria contributed to the decision-making process and to what extent. At this point we would like to repeat that ethical values might change over time. Therefore we need an agile automated system which can receive updates about the defined ethical principles.

Our Approach

Before an ethics check can take place, it must be defined what ethically correct means. Ethics is not universal. Ethics is something that is constantly evolving and is strongly influenced by culture. Together with our research partners from the University of Fribourg and from the Lucerne University of Applied Sciences and Arts we define ethical standards. The German-speaking European region was chosen as a first region. We call this region *smart region*, since we want to develop our standards based on the ideas of *smart cities*, which typically integrate the citizens and their environment into the process of idea-generation and development. With the help of the different research methods: literature reviews, surveys, expert interviews and focus groups, we will find out ethical standards for chat- and voicebots and give first approaches how these can also be measured. Opinions and knowledge of experts from various disciplines will be included in our research. These primarily include, psychology, data protection, computer science, digitization, education, data science, marketing, business. We need to consider that ethical values may change. Therefore, a mechanism of continuously improvements must be integrated. In further iterations, concrete methods or questions can then be developed to find out how ethically correct a bot behaves. or, in other words, to what degree a bot is still ethical and when its behaviour can no longer be called ethical. Providers, mostly companies that use bots for their customers and employees will then even receive suggestions on how they can develop their bots in a more ethically correct way. At this point, we probably have the challenge that many conversational AI projects are focused on specific use cases. So, we have to overcome the challenge of measuring ethical correctness despite these limitations, and possibly define rules for doing so. In the first phase, the ethics check will be carried out by humans. Humans chat as with the bot, ask the relevant questions, note down the answers and then evaluate them using a previously defined evaluation grid. While developing our ethic test, we keep attention to the work of Pangaro (incl. Pask [8]) on conversation theory to conceive a model allowing the learning, and thus the co-adaptation or co-evolution of interacting humans and/or machines, through conversation. Since nowadays AI often lacks of feedback loops, we want to integrate conversations which are about feedback loops on different levels managing the *how?* and the *why?*. In addition, conversation loops are elaborated to become a small data learning technique that brings many advantages such as ecological, economical, and privacy-preserving computing. At the point we would like to mention, that we only use the Turing test as inspiration in general. There are hardly any connections to the practical use. Here we rather focus on Pask [12] who describes intelligence as a property that is ascribed by an external observer to a conversation between participants where their dialogue must manifest understanding. He points out that each italicized word in this sentence requires careful attention. Furthermore, in an intelligent system theme, concept and memory are an important component [8].

Our final goal (until now) is that the chatting and the ethics check will be carried out by a chatbot, so that the entire ethics check can be fully automated in the long term. As soon as an AI must evaluate another bot, we will resort to the approaches of fuzzy logic and computing with words. Such fuzzy systems can deal with fuzzy data and are therefore very suitable when it comes to characterising the expressions of people or bots or testing their ethical maturity (i.e., fuzzy ethicity). Researchers who apply *fuzziness* not only classify words into categories, but also consider their position within the category [9]. Additionally, computing with words is a system of computation, based on fuzzy logic, in which the objects of computation are predominantly words, phrases and propositions drawn from a natural language, as we have in our chat conversations [9].

Current Status and Next Steps

The project is still at the early stage of development. For our work we use the method of synthetic modelling according to Kaufmann and Portmann [11]. We are currently in the analysis phase. With the help of a literature review, we are defining the first ethical guidelines for conversational Ais. In the next step, we will compare these with experts from the disciplines mentioned before and adapt them or develop new artefacts within the framework of the synthesis (cf. Kaufmann et al.). According to Kaufmann and Portmann [11], the approach of synthetic modelling with its creative parts is necessary so that new frameworks and technologies can be developed. Researchers are therefore encouraged to become creative and develop new artefacts and then test and optimize them again.

References

[1] Maedche, A. et al. (2019). AI-Based Digital Assistants. Business & Information Systems Engineering (61:4), 535– 544

[2] McLean, G. et al. (2019). Hey Alexa ... Examine the Variables Influencing the Use of Artificial Intelligent In-Home Voice Assistants. Computers in Human Behavior (99:April), 28– 37

[3] Dale, R. (2016). “The return of the Chatbots,” Natural Language Engineering (22:5). 811–817.

[4] Turing, A. M. The imitation game. Mind, 59(236):433–460,1950

[5] Teran L. et al. (2021). A Literature Review on Digital Ethics from a Humanistic and Sustainable Perspective, 14th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2021), Athens, Greece

[6] Alt, R. et al. (2021): Life Engineering. Bus Inf Syst Eng 63, 191–205. https://doi.org/10.1007/s12599-020-00680-x

[7] Floridi L (2018) Soft Ethics and the Governance of the Digital. Philos. Technol. 31, 1–8, https://doi.org/10.1007/s13347-018-0303-9

[8] Pangaro P. The Architecture of Conversation Theory, 1989.

[9] Zadeh, L. (1988). Fuzzy Logic. IEEE Computer, vol. 21, nr. 4, 83-93

[10] Portmann, E., D’Onofrio, S. (2022). Computational Ethics: Ein ethischer Lösungsansatz für das KI-Zeitalter. HMD 59 (2).

[11] Kaufmann, M. and Portmann E. (2017). Versuch einer Modellierung der Erkenntnispraxis mit Informationssystemen. Wirtschaftsinformatik in Theorie und Praxis, 73-83.

[12] Pask G. (1976): Aspects of Machine Intelligence. In: Nicholas Negroponte: Soft Architecture Machines, MIT Press, MA.