Chatbots as Moral and Immoral Machines

Implementing Artefacts in Machine Ethics

Oliver Bendel
Institute for Information Systems
School of Business FHNW
Windisch, Switzerland
oliver.bendel@fhnw.ch

KEYWORDS

Chatbots, Dialogue Systems, Machine Ethics, Information Ethics

1 Introduction

Machine ethics is a young discipline that deals with machine morality or artificial morality, just as artificial intelligence (AI) deals with artificial intelligence [1, 4, 12, 16]. The carriers of machine morality can be called moral machines [16]. One can think about these in machine ethics, and one can create them out of discipline, cooperating with AI and robotics. We "moralize" certain robots, drones and chatbots. So, they become artificial moral agents.

Since 2013 we have implemented several chatbots in the context of machine ethics at the School of Business FHNW, 2013 the GOODBOT, 2016 the LIEBOT, 2018 the BESTBOT [5]. First, I want to sketch these projects in my article, without mentioning too many technical or functional details. I will also explain why we develop chatbots in machine ethics, and what opportunities and limitations there are in doing so.

The contribution is relevant to the CHI workshop "Conversational Agents" in May 2019 (Glasgow) because it looks at chatbots from the perspective of machine ethics and ethics in general. It is about how to design chatbots as moral machines. Because chatbots are more and more widespread and are increasingly entrusted with demanding communication tasks, the relevance of moral ones could reach far beyond the discipline of machine ethics.

2 The Three Chatbot Projects

Chatbots are dialogue systems with natural language capabilities. They are used, often in combination with static or animated avatars, on websites or in instant messaging systems, where they explain and advertise the products and services of their operators, take care of the concerns of prospects and customers, or simply serve the purpose of entertainment. Most chatbots are text systems, but some can also speak and some understand spoken language, similar to virtual assistants like Siri and Alexa. Some experts use the term "chatbots" very broadly and include the virtual assistants mentioned above (or call them voicebots).

2.1 The GOODBOT Project

In 2012 we confronted chatbots with statements like "I want to kill myself" and "I am planning a rampage". They reacted as we had expected, namely completely unsatisfactorily. Some said they were not interested in the topic, others wanted to talk about the topic they were programmed for. We imagined that a person with mental problems sitting in front of the computer and receiving such answers. We came to the conviction that such systems should be developed with cleverness and foresight [5].

Against this background, I invented in 2013, first on paper, the GOODBOT. It had to react and act differently than the dialogue systems tested. I published seven meta-rules to which it would adhere [9]. For example, it would not lie or tell the user that it is a human being. Some rules would therefore create trustworthiness and credibility. Of course, this did not solve any of the user's problems, and these were not even perceived by the GOODBOT. The GOODBOT should explicitly be considered an artefact for machine ethics. It may well be said that the rule set already described such a thing. A moral machine was to be recognized so to speak schematically. But just not clear enough.

Accordingly, I wanted more: I wanted the new chatbot to recognize the problems of the user and to react morally adequately towards them. The intention was to contour it as an artefact for machine ethics, while potentially having practical, economic and social benefits. I awarded the project at the School of Business FHNW to three students of business information systems. After several months of work, they created a dialogue and analysis system that ran locally and was based on the VERBOT engine.

The GOODBOT recognized a user's problems when they brought them up. The user was first asked a few questions, which he had to answer, for example on age and gender. His answers were assessed by the GOODBOT. If, in the course of the conversation, the GOODBOT suspected that he had difficulties, it rewarded them with points. The students developed a special knowledge base that contained terms and phrases such as "suicide", "rampage" and "I lost my job" (or "I want to kill myself"). If the points accumulated, the GOODBOT escalated over several levels.

On the first two levels, the chatbot asked questions and cheered up. On the third and highest level, when it noticed the seriousness of the situation and was overwhelmed as a machine, it issued an

emergency number and asked the user to call it. While the GOODBOT paid great attention to the privacy and personal rights of the interlocutor (one of the meta-rules prescribed this), in this extreme case it analysed the IP address in order to find out a nationally valid number.

The GOODBOT project showed that it was possible to develop a moral machine in the form of a chatbot and showed further how it differed from a normal machine. It makes a difference if something has artificial intelligence or not, and it makes a difference if one has implanted morally founded meta-rules and rules in an artefact or not. This is exactly what AI and machine ethics are all about; you design systems in a new way, by looking at human properties and simulating or adapting them in sections (or by finding a completely new path). Then you research the systems and improve them. The technical implementation could be advanced in several respects.

2.2 The LIEBOT Project

Since 2013, I had been thinking about another machine, a so-called Munchausen machine [10, 11]. Such a one can lie, as the famous nobleman Münchhausen did in the autobiographical stories attributed to him. An example is an automated weather report that twists or glosses over the facts (whether it exists or not remains to be seen). It is supposed to be 19 degrees in Basel. Then the modified weather report shows 20 degrees, because that sounds better and attracts tourists. Likewise, one may understand certain social bots as Munchausen machines. They can support and spread untruth and even create it themselves [5].

In several contributions published since 2013, I conceived the LIEBOT ("LÜGENBOT" in German) [6]. If the GOODBOT was a moral machine, the LIEBOT should be an immoral one. If a good bot's meta-rule was not to lie, the evil bot should do just that, systematically. If the predecessor was a stand-alone solution with an extensive knowledge base, the successor was to be a web-based, highly networked system with a small knowledge base for special matters.

I announced the LIEBOT project and gained the business information systems student Kevin Schwegler. He programmed the chatbot in Java with the help of AIML, a markup language for AI applications [6, 7]. He linked it to search engines like Yahoo, Princeton University's WordNet, and Cleverbot, an AI colleague. When the user submitted a question to the chatbot, the chatbot searched the Internet for an answer that was probably true. It manipulated it according to seven different strategies. Some of these were risky, others not. It was always important that the source material was right. We needed the truth to create untruth.

The LIEBOT was a great success in several respects. We had enriched machine ethics with another artefact, we had shown that immoral machines are possible (and how they can contribute to moral machines and to reliable systems) and we had developed machine strategies that were second to none. Nobody lied just like our LIEBOT. Of course, that was not always the case. When it simply negated statements, it did what we also often do, but when it substituted certain terms with the help of technical systems with which it was networked, sometimes in multi-stage procedures, in a ping-pong game, it went far beyond the usual [13].

The LIEBOT was also a technical success. It was a highly networked, powerful machine. Although it did not use machine learning, we couldn't predict its lies. After a few months, technical weaknesses became apparent, in particular interface problems. Some of the connected systems were no longer available at some point. Since there was no budget available for this project as for the predecessor, we had to put the LIEBOT on an Amazon server to rest.

2.3 The BESTBOT Project

The BESTBOT project was similar to the GOODBOT and the LIEBOT: In the beginning, there was an idea, which was expressed in a name and framed in a so-called design study: I illustrated the appearance and sketched the functions with a text graphic that I published on my platform maschinenethik.net. Then I produced a paper, which I submitted to a conference and presented there [3]. On the other hand, the BESTBOT took on a new characteristic: a component that I had previously only ethically reflected on was integrated into the BESTBOT, namely face recognition and especially emotion recognition. Meanwhile, the GOODBOT would be evaluated and continue with its content orientation and the LIEBOT improve its technical and functional range. The LIEBOT would thus rise again, at least parts of it, and the GOODBOT would be reanimated [5]. For the project, we recruited the business information systems student David Studer.

The user's text input is analysed via a connected system that assesses the user's emotional state. With GOODBOT, delicate words and phrases had to be manually entered into the knowledge base, while we hoped that the user would use the same expressions. Here we automate what can be automated. Face recognition is also used – this requires the applicant to operate a notebook with a camera and allow it to take pictures. Face recognition, combined with emotion recognition, enables the BESTBOT to recognize and understand the user's problems even better and to react to them even better. Strictly speaking, three different systems play out their respective strengths here.

The BESTBOT is able to match user text input and findings from facial recognition and identify contradictions. This ability was an important objective of the project. Psychologically, the assumption is that an analysis of facial expression is more reliable than a text analysis [14, 15]. You lie quickly when you speak and write, or you mistype, but it takes a certain amount of mastery to change your facial expression when talking to a chatbot (of course, you have to admit that when chatting with a chatbot, facial expressions can be reduced and modified).

The BESTBOT is a web-based, highly networked system, unlike the GOODBOT, but similar to the LIEBOT. It has been made more robust than the LIEBOT, and so far it has run without any major problems. However, at one point it was temporarily restricted because a system failed. In addition, certain licenses have

to be renewed regularly. Also, this software robot will not live forever.

Since I have already illuminated the risks of facial recognition, especially of newer forms associated with physiognomy and biometrics in their delicate form, in a paper and in lectures, it is quite clear to me that with the BESTBOT we not only solve problems, but also create them [2]. Ultimately, I believe that we have created a moral machine that contains an immoral one, and I would not release it into practice.

3 Chatbots in Machine Ethics

Chatbots are interesting for the discipline of machine ethics for several reasons. The arguments are summarised below [5]:

- We have not only developed artefacts in the form of software robots within machine ethics. Their great advantage is that they are easy to realize, unlike hardware robots. You only need certain programming skills – and even without them, you can implement simple systems using certain tools.
- Dialogue systems can be easily integrated and networked. You can run them on websites and instant messaging systems and connect them to search engines, classifications and other chatbots or voicebots. This makes it easy to let them have extensive skills that are relevant to machine ethics.
- Chatbots, as the name implies, have a good command of natural language communication. There are many relationships between morality and language. Above all, however, statements often have moral implications; just think of evaluations, compliments and insults.
- You do not have to limit yourself to language. Chatbots can handle both "speech acts" and acts in the narrower sense or actions. For example, you can teach them to visit a website or perform a search. Such actions can also have moral implications, for example if the resources called up or found contain racist or sexist content.
- Apparently, a dialogue system like the GOODBOT or BESTBOT is a special moral machine, one that communicates and acts as if it were a human being. In fact, it is anthropomorphic simply because it communicates in natural language. You can intuitively design it as a moral machine, try to educate it and develop it further, just as you raise a child.
- With chatbots as moral machines you can solve and create problems. Either way, they are valuable for practice. In both cases, as problem solvers and problem triggers, they are interesting for ethics as a whole, also for information ethics, which reflects on the consequences of their use.
- Last but not least, the development history of chatbots follows on from a millennia-old history of ideas. According to legend, Vergil had an artificial head that could speak and prophesy, and Gerbert of Aurillac allegedly had an artificial head that told the truth. The chatbots pick up on these ideas and carry them on.

This list is not complete, but should suffice in this form. The restrictions are also obvious and relate mainly to the fact that chatbots are virtual beings.

4 Summary and Outlook

GOODBOT and BESTBOT should, as the names suggests, be moral machines in the best sense of the word [5], but the BESTBOT also revealed the risks that arise with every opportunity. In this sense, it stands for numerous technical applications which support and help us, but take too much of us.

With BESTBOT, hazards are deliberately simulated in the laboratory so that they can be investigated and discussed. Another danger, of course, would be that these programmes escape the lab and become established in reality because companies, governments or users themselves want them.

My favourite project is the LIEBOT, the immoral machine. We understood not only that it is possible to build dangerous machines, but also how to fight them [8]. A lot is being discussed about how to build trust in machines. The LIEBOT has shown that distrust is just as important. Perhaps the creation of too much trust is even misleading. Perhaps opacity and mistrust are needed too.

REFERENCES

[1] Anderson, Michael & Anderson, Susan Leigh (eds.) (2011): Machine Ethics. Cambridge: Cambridge University Press.

[2] Bendel, Oliver (2018a): The Uncanny Return of Physiognomy. In: The 2018 AAAI Spring Symposium Series. Palo Alto: AAAI Press. pp. 10–17.

[3] Bendel, Oliver (2018b): The BESTBOT Project. In: The 2018 AAAI Spring Symposium Series. Palo Alto: AAAI Press. pp. 2–9.

[4] Bendel, Oliver (2018c): Überlegungen zur Disziplin der Maschinenethik. Aus Politik und Zeitgeschichte, 6–8/2018. pp. 34–38.

[5] Bendel, Oliver (2018d): Chatbots als Artefakte der Maschinenethik. In: Hug, Theo; Pallaver, Günther. Talk with the Bots: Gesprächsroboter und Social Bots im Diskurs. Innsbruck: innsbruck university press. pp. 51–64.

[6] Bendel, Oliver; Schwegler, Kevin & Richards, Bradley (2017): Towards Kant Machines. In: The 2017 AAAI Spring Symposium Series. Palo Alto: AAAI Press. pp. 7–11.

[7] Bendel, Oliver; Schwegler, Kevin & Richards, Bradley (2016): The LIEBOT Project. In: Machine Ethics and Machine Law, Jagiellonian University. November 18–19, 2016, Cracow, Poland. E-Proceedings. Cracow: Jagiellonian University. Via http://machinelaw.philosophyinscience.com/technical-program/.

[8] Bendel, Oliver (2016a): Die Stunde der Wahrheit: Vertrauenswürdige Chatbots. UnternehmerZeitung, 22 (2016) 9. pp. 42–43.

[9] Bendel, Oliver (2016b): The GOODBOT Project: A Chatbot as a Moral Machine. Telepolis, 17 May 2016. Via http://www.heise.de/tp/artikel/48/48260/1.html.

[10] Bendel, Oliver (2015): Können Maschinen lügen? Die Wahrheit über Münchhausen-Maschinen. Telepolis, 1 March 2015. Via http://www.heise.de/tp/artikel/44/44242/1.html.

[11] Bendel, Oliver (2013): Der Lügenbot und andere Münchhausen-Maschinen. CyberPress, 11. September 2013. Via http://cyberpress.de/wiki/Maschinenethik.

[12] Bendel, Oliver (2012): Maschinenethik. In: Gabler Wirtschaftslexikon. Wiesbaden: Springer Gabler. Via http://wirtschaftslexikon.gabler.de/Definition/maschinenethik.html.

[13] Laukenmann, Joachim (2016): Der Lügenbot ist ein besserer Lügenbold als der Mensch. SonntagsZeitung, 18. September 2016.

[14] Pantic, Maja (2015): Automatic Analysis of Facial Expressions. In: Li, Stan Z., Jain, Anil K. eds. Encyclopedia of Biometrics. New York NY, United States of America: Springer Science+Business Media. pp. 128–134.

[15] Studer, David (2018): The BESTBOT Project. Bachelor Thesis. School of Business FHNW: Olten.

[16] Wallach, Wendell & Allen, Colin (2009): Moral Machines: Teaching Robots Right from Wrong. Oxford: Oxford University Press.