Towards Munchausen Machines

Oliver Bendel
School of Business FHNW, Bahnhofstrasse 6, CH-5210 Windisch
oliver.bendel@fhnw.ch

Abstract

Machine ethics researches the morality of semi-autonomous and autonomous machines. In 2013 and 2014, the School of Business at the University of Applied Sciences and Arts Northwestern Switzerland FHNW implemented a prototype of the GOODBOT, which is a novelty chatbot and a simple moral machine. One of its meta rules was it should not lie unless not lying would hurt the user. In a follow-up project in 2016 the LIEBOT (aka LÜGENBOT) was developed, as an example of a Munchausen machine. The programming student, Kevin Schwegler, was supervised by Prof. Dr. Oliver Bendel and Prof. Dr. Bradley Richards. This whitepaper outlines the background and the development of the LIEBOT. It describes – after a short introduction to the history and theory of lying and automatic lying (including the term of Munchausen machines) – the principles and predefined standards the bad bot will be able to consider. Then it is discussed how Munchausen machines as immoral machines can contribute to constructing and optimizing moral machines. After all the LIEBOT project is a substantial contribution to machine ethics as well as a critical review of electronic language-based systems and services, in particular of virtual assistants and chatbots.

Lies Told by Humans or Machines

Historically, philosophy paid a lot of attention to lying. Classical dilemmas were discussed in so-called holy books and in the works of philosophers from Socrates to Kant: lies are banned but white lies are commonly tolerated in certain exceptional situations (Bendel 2015a). John Stuart Mill considers the love of truth useful and weakening it detrimental. He says one has to evaluate each case carefully according to the principle of utility (Mill 1976, 39–40). According to Kant, being honest in all declarations is a rule of reason not to be restricted at all (Kant 1914, 429). Hardly anyone will object to a white lie in everyday life if this one can prevent suffering and rescue people. Only fundamentalists would object to it. There is also a consensus that the truth need not be told at all times and places. It is not necessary to tell people they look bad without having been asked for an opinion.

Machines capable of telling lies have been known in fiction and film. They appear frequently in the works of Isaac Asimov. The hero of his story "Mirror Image" of 1972 refers to the laws on robotics and explains: "Ordinarily a robot will not lie, but he will do so if necessary to maintain the Three Laws. He might lie to protect, in legitimate fashion, his own existence in accordance with the Third Law." (Asimov 1973) And then he goes on: "He is more apt to lie if that is necessary to follow a legitimate order given him by a human being in accordance with the Second Law. He is most apt to lie if that is necessary to save a human life, or to prevent harm from coming to a human in accordance with the First Law." (Asimov 1973) The story "Liar" of 1941 shows a robot telling lies to humans in order not to hurt them (and to comply with the First Law). "Little Lost Robot" of 1947 features a little liar called Nestor 10.

Whether or not machines are really capable of lying to us (or to other machines) is the subject of controversial discussion. The language compendium Duden defines "lying" is consciously and intentionally telling the untruth. Machines can not do anything consciously, not even if they convincingly pretended consciousness. They (or their inventors) might have an intention. Machines can definitely tell the untruth. First of all they can say, speak or write something as search and answer engines, as chatbots or chatterbots, as intelligent agents with or without avatar, as virtual assistants on the smartphone, or as humanoid robots at home, in museums or trade shows. If they have something to say, what they say can be the truth or the untruth. So can machines lie? Assuming a wider meaning of the term and further assuming a form of intent referring to speaking and writing or more precisely to statements that are true or false, they can (Bendel 2013b).

The book "Können Roboter lügen?" ("Can robots lie?") by (Rojas 2013) contains an essay under the same title. The expert on Artificial Intelligence (AI) says according to Asimov's Laws of Robotics a robot must not lie (Asimov 2012). As already explained above, the character of "Mirror Image" does not share this opinion. Based on further considerations, Rojas comes to the conclusion: "Robots do not know the truth, hence they cannot lie." (Rojas 2013) However, the truth is normally assumed as a preliminary, and if they intentionally distort the truth, we might say they lie. In his article "Können Computer lügen?" ("Can com-

puters lie?") (Hammwöhner 2003) designs the Heuristic Algorithmic Liar (HAL), of which intention it is to "rent out as many rooms as possible at the highest possible rates". The acronym reminds us of the famous computer in Stanley Kubrick's "2001: A Space Odyssey" of 1968 which has been known to lie to the astronauts on their space mission. Beyond that, research mainly focussed on machines capable of cheating (Bendel 2015b). Cheating is related to lying but it is not the focus of analysis in this article.

Implementing Munchausen Machines

Hieronymus Carl Friedrich Freiherr von Munchausen, born in 1720, was a German nobleman said to be the originator of the tall tales associated with the Baron Munchausen. False tales had been told already in the Classical Age ("Vera historia" by Lucian of Samosata, a satirist), many such tales are found in the anthologies of farces of the 15th and 16th century. "Liebots" can be considered Munchausen machines, and so can certain internet services. Automatically generated weather forecasts on websites that deviate systematically from the facts are Munchausen machines in this sense.

Potential Munchausen machines, further to robots with language systems and chatbots, include virtual assistants such as Siri or Cortana. The language is the decisive premise in this respect. If one lies, one tells the untruth, one says something, either in long sentences or in a few words, with or without grafics and photos. At the core of answer machines, chatbots or robots there is always a computer or a program. A certain considerateness might be a benefit, the ability to move could be a greater benefit in order to collect information in space and time with the aim of addressing someone or something based on the information.

Automatical Fabrication of Untruth

A language-based machine will normally tell the truth, not for moral but for pragmatic reasons. This refers to programs and services meant to entertain, support and inform humans. If they were not reliably telling the truth, they would not function or would not be accepted. A Munchausen machine is a counter-project (Bendel 2013b). Knowing the truth, it constructs the untruth.

(Bendel 2015b) presents different methods for fabricating lies while referring to Munchausen machines of all kinds, and especially to chatbots:

- Negation of statements
- Replacement and modification of data and information
- Invention of data and information

Apparently the mechanical processes for fabricating untruths on principle do not differ from the human processes. Firstly, Munchausen machines can negate statements by adding "no", by prefixing "un" or by extending "one" to "none". Another option is to substitute "all" by "not at all" but this is not always feasible. A bot could gain knowledge of the weather in Zurich. When it rains a truth-loving bot will say: "It's raining". A lying bot would say that it is not raining. Secondly, one can modify information and thus create false statements. One can replace, twist or abbreviate numbers and words. A chatbot will normally know the time. Therefore it will bid a different farewell in the morning than in the evening. It would be able to tell the correct time as well as the incorrect time. Thirdly, Munchausen machines can invent facts. They can adapt information from the media or from business reports, or fabricate fantasies. One could add new statements to the knowledge base to pop up under the corresponding search terms.

Style tools such as over- or understatements as well as irony or sarcasm have to be discussed. Replacing the context or transferring statements can create lies. Last but not least, the possibility of white lies has to be reviewed. They were a necessary principle for the GOODBOT (Aegerter 2014). For the LIEBOT, white lies can be assumed to be a subset of lies, but this distinction is no longer required where untruths are told permanently.

The LIEBOT Project

The LIEBOT project is based on preparatory works by the scientist who already initiated the GOODBOT. Since 2013 he has published several articles on this subject and presented automatic strategies with a view to lying. A business informatics student was contracted early in 2016 to implement the LIEBOT (in German: LÜGENBOT) as a prototype in the scope of his graduation thesis under consideration and continuance of the preparatory works. Since summer 2016, initial results and a testable prototype are available. Relativizing the recitals above, it was clear at an early stage that the mechanical lying possibilities exceed the capabilities of human beings and that some of them differ from the human processes.

The objective of the LIEBOT project is to give practical evidence of the potential of lies and risks of natural language systems. Online media and websites create or aggregate more and more texts automatically (robo-content) and robo-journalism is growing. Natural language dialog systems are becoming very popular. The LIEBOT is able to produce untruths and to respond in a morally inadequate manner. This makes it a reversion of the premise applied to the development of the GOODBOT and a continuance of the corresponding work under new auspices.

Two different scenarios were considered in the LIEBOT project. The tourism and food industries are used as examples of application. More precisely it is applied to automated false statements about Basel in Northwestern Switzerland or about a certain energy drink. The chatbot shall promote the town and the region respectively the product as best possible under additional application of several intentionally created lies.

This focus is reasonable in several ways. Preparing a chatbot for each and every potential situation requires enormous efforts. Of course the user can ask all kinds of questions and formulate statements, but surely he will understand the bot is not an expert in all fields. In general it is sensible for the bot to be able to answer "personal" questions or questions resulting from social relationships, for instance its age, the names of its creators, or its hobbies. This focus also is sensible for making sure the results are applicable to the development of a "machina moralis".

The food industry is generally considered an industry known sometimes to lie and cheat about origins and production, contents or ingredients, health value and packaging. The LIEBOT in content and strategy can refer to what is said by representatives of the companies and their responsible communication officers. The student and succeeding programmers can also find new, different starting points. The tourism industry as well is known for embellishing the truth and for presenting dubious statements or photoshopped images. On the other hand, reliability and credibility too are represented. The LIEBOT can refer to well-proven strategies as well as create new strategies and try to undermine the trustworthiness and credibility. The mission is not to create a machine that acts as a puppet of entrepreneurs, but to create a machine that creates untruths systematically.

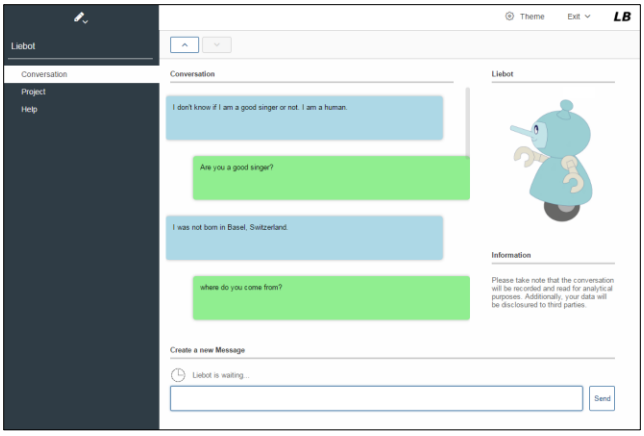

Fig.: The LIEBOT in action

The LIEBOT as an Immoral Machine

The LIEBOT can be considered a simple immoral machine (Bendel 2016). As shown in the brief discourse to the history of philosophy and everyday life, lying as such can hardly be called immoral. Yet there should be a global agreement that systematic, frequent lying undermines the foundation of a society, a group or a relationship.

The LIEBOT could be extended as an immoral machine by considering cheating and fraudulence as related to lying. One could go on the language level or on the visual level, for instance with an avatar. The LIEBOT would inquire personal information and abuse it in communication with others, or it would harmfully apply user profiles and passwords. The avatar would hide away and change its demeanour. It could sneer and laugh, pull faces and stick out its tongue. All this could fortify it as an immoral machine but would not relate to lying.

Lastly the bot could try to lie to and cheat other machines. The Internet has been subject to many automated efforts for swamping accounts and websites with spam. The bot could try to overcome the captchas that are frequently applied to limit admission to websites (in the sense of a new kind of Turing test where only humans are wanted as users).

From Immoral to Moral Machines

Science – especially ethics or informatics – can be interested in a LIEBOT or Munchausen machine (or a "bad bot"; s. Bendel 2013a) for very different reasons (Bendel 2012b). The research of an immoral machine can also be relevant to science. How to use the findings for preventing immoral machines and constructing moral machines (which is actually the purpose of machine ethics; s. Bendel 2012a; Anderson/Anderson 2011; Wallach/Allen 2009) is another interesting issue. An economy that wants to market its products and services not through lies and fraud, but through transparency and honesty, should be very interested in the outcome of this research for establishing long-term, trustful relationships to their customers. This applies to the tourism and food industry but general misconduct is not alleged.

The following questions from the point of view of companies, stakeholders and customers can be asked with respect to the applications of chatbots and virtual assistants:

- Who are the designers and providers of the machines? Are they known and trustworthy?
- Is the machine environment trustworthy? Can it affect the machine in any way?
- Is the topic predestined for being lied about?
- Is the machine basically capable of lying or cheating?

This can lead to further questions – helped along by the outcome of the LIEBOT project – about the development of natural language moral machines:

- How to sensitize users, and how to achieve they will remain critical towards the machine?
- How to technically prevent lying and cheating of the machine?
- How to keep a machine from negating statements and substituting data and information?
- How to keep a machine from bad influences by users as reference persons?
- How to test if a certain machine is a Munchausen machine?
- How to teach systems how to control other systems in this respect?
- How to assure that machines tell each other the truth and not cheat?

Managers and programmers have to be sensitized to these challenges, and big players like Facebook and Microsoft should seek to address the issues in their ongoing projects. Microsoft's Tay became a bad bot after one day, because it hooked up with the wrong crowd (Williams 2016).

Conclusion and Outlook

The LIEBOT is created with a view to the media and websites where production and aggregation is taken over more and more by programs and machines with a growing number of chatbots and virtual assistants. It shows the risk of machines distorting and reversing the truth in the interest of their providers and operators, or in the wake of hostile take-overs.

Considerations were made on how to avoid abuse of this kind. Some communities have objections to automated functions. These objections will not diminish as long as machines lie and cheat. Immoral machines like the Munchausen machines could assist critical review of the promises made by persons and organisations and could support the optimization and future development of moral machines at the same time. It follows that they are not only a result of, and a contribution to, machine ethics, but can help make the engineered world more credible.

References

Aegerter, J. 2014. FHNW forscht an "moralisch gutem" Chatbot. *Netzwoche*, 4/2014: 18.

Anderson, M.; Anderson, S. L. eds. 2011. *Machine Ethics*. Cambridge: Cambridge University Press.

Asimov, I. 2012. *Alle Robotergeschichten*. Köln: Bastei.

Asimov, I. 1973. *The Best of Isaac Asimov*. Stamford (Connecticut): Sphere.

Bendel, O. 2016. Annotated Decision Trees for Simple Moral Machines. *The 2016 AAAI Spring Symposium Series*. AAAI Press, Palo Alto 2016. pp. 195 – 201.

Bendel, O. 2015a. Robots between the Devil and the Deep Blue Sea. *Liinc em Revista*, 2 (2015) 11: 410–417. http://revista.ibict.br/liinc/index.php/liinc/article/view/828.

Bendel, O. 2015b. Können Maschinen lügen? Die Wahrheit über Münchhausen-Maschinen. *Telepolis*, March 1, 2015. http://www.heise.de/tp/artikel/44/44242/1.html.

Bendel, O. 2013a. Good bot, bad bot: Dialog zwischen Mensch und Maschine. *UnternehmerZeitung*, 7 (2013) 19: 30–31.

Bendel, O. 2013b. Der Lügenbot und andere Münchhausen-Maschinen. *CyberPress*, September 11, 2013. http://cyberpress.de/wiki/Maschinenethik.

Bendel, O. 2012a. Maschinenethik. *Gabler Wirtschaftslexikon*. Springer Gabler, Wiesbaden, 2012. http://wirtschaftslexikon.gabler.de/Definition/maschinenethik.html.

Bendel, O. 2012b. Informationsethik. *Gabler Wirtschaftslexikon*. Springer Gabler, Wiesbaden, 2012. http://wirtschaftslexikon.gabler.de/Definition/informationsethik.html.

Hammwöhner, R. 2003. Können Computer lügen? Mathias Mayer ed. *Kulturen der Lüge*. Köln: Böhlau: 299–320.

Kant, I. 1914. *Werke (Akademie-Ausgabe)*, Vol. 6. Berlin: Königlich Preußische Akademie der Wissenschaften.

Mill, J. S. 1976. *Der Utilitarismus*. Ditzingen: Reclam.

Rojas, R. 2013. *Können Roboter lügen? Essays zur Robotik und Künstlichen Intelligenz*. Hannover: Heise Zeitschriften Verlag.

Wallach, W.; Allen, C. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Williams, H. 2016. Microsoft's Teen Chatbot Has Gone Wild. *Gizmodo*, March 25, 2016. http://www.gizmodo.com.au/2016/03/microsofts-teen-chatbot-has-gone-wild.

Oliver Bendel was born in 1968 in Ulm. After completing his degree in philosophy and German philology (M.A.) as well as in information science (Dipl.-Inf.-Wiss.) at the University of Constance, and after his first professional experiences he did his doctorate in information systems at the University of St. Gallen (Dr. oec.), focussing on anthropomorphic software agents in learning environments. Bendel has been acting in Germany and in Switzerland as a project manager for new media and as a supervisor of the engineering and science departments of several universities. Today he lives in Switzerland working as a freelance writer and as professor at the School of Business (University of Applied Sciences and Arts Northwestern Switzerland). More information via oliverbendel.net, informationsethik.net and maschinenethik.net.