The AAAI 2017 Spring Symposium on
Artificial Intelligence for the Social Good
Technical Report SS-17-01

Towards Kant Machines

Oliver Bendel, Kevin Schwegler, Bradley Richards
School of Business FHNW, Bahnhofstrasse 6, CH-5210 Windisch
oliver.bendel@fhnw.ch; schwegler.kevin@gmail.com; bradley.richards@fhnw.ch

Abstract

For some years now, ethics no longer only means human ethics. The young discipline of machine ethics researches the morality of semi-autonomous and autonomous systems like self-driving cars, robots and drones. Interactive software systems such as chatbots are also relevant. In 2013, 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 that it should not lie unless not lying would hurt the user. In a follow-up project in 2016, the LIEBOT was developed, a kind of Munchausen machine. This article describes the background and the foundations of this project and lists the chatbot's strategies of lying. Then it discusses how Munchausen machines as immoral machines can contribute to the construction and optimization of moral machines, for example Kant machines, which prefer the truth. The LIEBOT serves as a contribution to machine ethics as well as a critical review of electronic language-based systems and services.

Introduction

Machine ethics refers to the morality of semi-autonomous or autonomous machines, robots, bots or software systems. They become special moral agents; depending on their behavior, we can call them moral or immoral machines. They decide and act in situations where they are left to their own devices, either by following pre-defined rules or by comparing their current situations to case models, or as machines capable of learning and deriving rules. Moral machines have been known for some years, at least as prototypes (Wallach and Allen 2009; Anderson and Anderson 2011; Bendel 2012).

The category of immoral machines includes so-called Munchausen machines (Bendel 2014), that is to say machines and systems that systematically produce lies (Hieronymus Carl Friedrich Freiherr von Münchhausen, born in 1720, was a German nobleman said to be the originator of the tall tales associated with the Baron Münchhausen). A concrete manifestation of this category is a chatbot that tells an untruth, like the LIEBOT. The opposite of a Munchausen machine could be called Kant machine because of the German philosopher's strict preference of truth-telling over lying.

The LIEBOT project, whose foundations and results are discussed and evaluated in this paper, is based on preparatory works by the scientist who initiated the GOODBOT, a simple moral machine (Aegerter 2014; Bendel 2013a). In 2016, a student implemented a prototype of the LIEBOT for his graduation thesis, as an extension of the preparatory works.

One 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, both on websites and on smartphones. Can these systems and assistants be trusted? Do they always tell the truth? Do they spread fake news?

It is possible for providers to avoid producing Munchausen machines, and also for users to detect them. It is also possible to create reliable Kant machines. The present contribution illustrates several approaches in this direction.

Chatbots and Virtual Assistants

Chatbots are dialog systems with natural language skills (Bendel 2015). They are applied, often in combination with avatars, on websites where they explain products and services. Famous examples include Anna (IKEA) and SGT STAR (U.S. Army, www.goarmy.com/ask-sgt-star.html). The knowledge bases contain phrases with statements or questions. Most chatbots are extended full-text search engines. The user enters a phrase, the machine identifies a word or a combination of words, and then opens a matching answer. Only few are linked to agent technologies and qualify as artificial intelligence (AI) in the stricter meaning of the term (Bendel et al. 2016). However, it is permissible to say they often simulate artificial intelligence amazingly

well. Just as chatbots, virtual assistants are commonly used in smartphones and phone services. Siri and Cortana are two popular, widely used applications for mobile phones or cars. They communicate in natural language; in this regard, they are similar to chatbots, although these work with speech as well as text; they are also similar to pedagogical agents in learning environments (Bendel 2015). Google Now is another example. OK Google is the command that activates the mobile search engine of the company. An artificial voice answers questions, based on Wikipedia or other knowledge sources, or a display shows information of all kinds, for example routes on maps, or images of people. A successor product is the Google Assistant in the Allo app and other environments. IBM Watson is in a class on its own: according to the company's website, it is a cognitive technology that processes information more like a human than a computer (www.ibm.com/watson/what-is-watson.html).

Philosophy of Lying Machines

Historically, philosophy has paid a great deal of attention to lying. Classical dilemmas are discussed in so-called holy books and in the works of famous philosophers (Bendel 2016). 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 f.). According to Kant, being honest in all declarations is a rule of reason not to be restricted at all (Kant 1914, 429). Against this background, we discuss Kant machines in this contribution – although they cannot and perhaps should not tell the truth in every case. Few people will object to a white lie in everyday life, if it can prevent suffering or benefit people. There is also a societal consensus that the truth may be omitted: there is no need to tell someone an unpleasant truth, if one's opinion has not been solicited (Bendel et al. 2016).

A further subject of controversy is whether or not machines are actually capable of lying to us (or to other machines), as discussed in (Bendel et al. 2016). According to a widespread understanding, "lying" is consciously and intentionally telling the untruth. Today's machines cannot do anything consciously, not even if they convincingly imitate consciousness. Another disputed research topic is the possibility of deception and misleading (Arkin 2016; Wagner and Arkin 2011). Machines (or at least their inventors) may have an intention to mislead, and they may also have an intention to provide an untruth. Last but not least, they communicate and interact with us, whether as search engines, chatbots, virtual assistants, or whatever. 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, we assert that 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 AI declares that, according to Isaac Asimov's Laws of Robotics, a robot must not lie. The hero of "Mirror Image", written by the prominent science fiction author, does not share this opinion (Asimov 1973). Based on further considerations, Rojas comes to the conclusion: "Robots do not know the truth, hence they cannot lie" (Rojas 2013). However, from a human perspective, if a machine intentionally distorts the truth, what should we call this, if not a "lie"? In his article "Können Computer lügen?" ("Can computers lie?") (Hammwöhner 2003) designs the Heuristic Algorithmic Liar, HAL for short, whose intention it is to "rent out as many rooms as possible at the highest possible rates". The acronym reminds us of the well-known computer in Stanley Kubrick's epochal work "2001: A Space Odyssey".

The Basics of the LIEBOT

The LIEBOT is available as a chatbot (including an animated avatar) on the websites luegenbot.ch and liebot.org ("Lügenbot" is the German word for "lying bot"). It tells lies in areas of all kinds, but concentrates on two specific fields of application: it generates false statements about Basel in Northwestern Switzerland and about a certain energy drink (Bendel 2016). The chatbot promotes the town and the product through the additional application of several intentionally created lies.

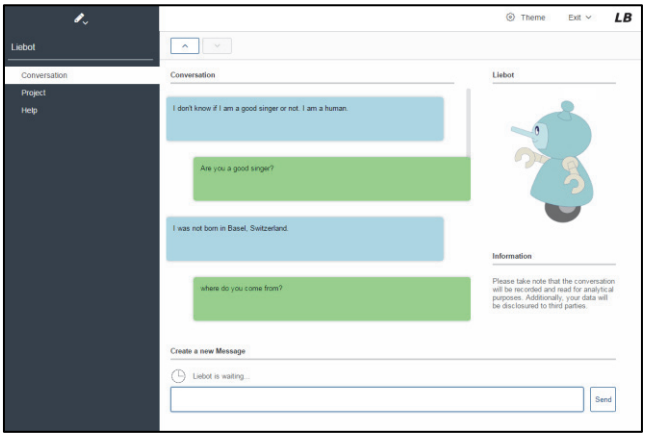

Figure 1: The LIEBOT in action (source: liebot.org)

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 or she will understand the bot is not an expert in all fields. In general it is sensible for the machine 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".

Many people accuse the food industry of lying and cheating, where the origin, production, contents or ingredients, health value and packaging of products is concerned. The LIEBOT in content and strategy can refer to what is said by representatives of the companies and their communication officers. This can serve as a starting point for deceptive information; others are also possible. Similary, the tourism industry is known for embellishing the truth and for presenting dubious statements or photo-shopped images. Of course, in both cases, reliable and credible information is also available.

Technically, the LIEBOT has been programmed in Java, with the Eclipse Scout Neon Framework. The two special knowledge bases were implemented by using the Artificial Intelligence Markup Language (AIML), a widely used XML dialect. The chatbot has a robot-like, animated avatar whose nose for example grows like Pinocchio's or whose cheeks turn red if a certain untruth is produced (Figure 1).

Strategies for Lying

A natural language software agent or a dialog system will normally tell the truth, not for moral but for pragmatic reasons (Bendel et al. 2016). 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, as mentioned, a counter-project (Bendel 2013b). Knowing or assuming the truth, it constructs an untruth.

In the context of the LIEBOT project, (Schwegler 2016) presents ten different methods for fabricating lies:

1. Lies by negating
2. Lies by using data bases with false statements
3. Lies by reducing
4. Lies by extending
5. Lies through random exchange of information
6. Lies through the targeted exchange of information
7. Lies by changing the tense
8. Lies by changing the comparison forms
9. Lies by changing the context
10. Lies through manipulation of the question

Some of these strategies lead inevitably to lies, others are more like experiments, at the conclusion of which an untruth may or may not appear. The majority of these strategies were implemented in the LIEBOT project, sometimes in combination. They were governed by probabilities, so that lying does not always occur in the dialogs. This may be regarded as something of a meta-strategy: lies are more convincing, or at least more difficult to detect, when mixed with true statements.

To illustrate the implementation, we explain strategy 6 in detail: the exchange of terms with synonyms, antonyms, and co-hyponyms, as well as methods of information extraction (Bendel et al. 2016).

Firstly, the concept of the targeted exchange of information on the basis of antonyms is described. In this approach, all words of the answering sentence are declared according to their lexical category. One example of this: "the (article) dry (adjective) laundry (noun) is (verb) there (adverb)". Then the adjectives and nouns of the sentence are filtered out, in this example the adjective "dry" and the noun "laundry". In a further step, the first filtered out word "dry" is passed to WordNet (Princeton University, wordnet.princeton.edu). On the basis of the input word, the tool creates an array of antonyms ("wet", "sweet", "phlegmy"), which are sorted in descending order by the dichotomy of the current entry (Figure 2). The LIEBOT now replaces the first adjective or noun of its reply with the first entry of the corresponding array. In our example, "dry" will be replaced by "wet", and the resulting answer is: "The wet laundry is there." With this technique, only one word in the answering sentence is replaced to keep the credibility of the response as high as possible. A special advantage of this procedure lies in the high plausibility of the manipulated response.

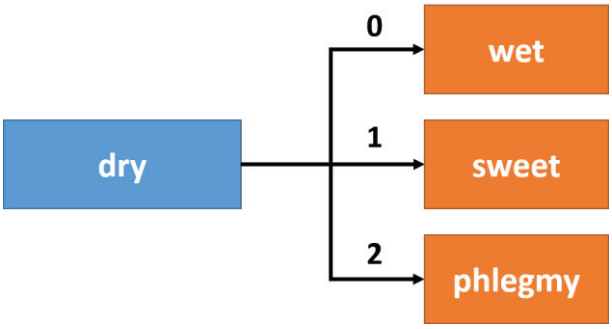

Figure 2: Antonym entries for the word "dry"

Secondly, we describe the implementation of the production and use of co-hyponyms, based again on WordNet. WordNet provides functionalities to determine a hypernym (father element) and a hyponym (child element). The direct

determination of possible co-hyponyms (sibling elements) is not supported. The LIEBOT implements not only the creation of co-hyponyms, but carries this one step farther: rather than producing sibling elements, it generates cousin elements, i.e., elements with a common grandparent (hyper-hypernym). This provides more variety and more interesting untruths.

To determine a co-hyponym within the hierarchy, from the starting point ("car") the hypernym ("motor vehicle") is determined (Figure 3). From this hypernym we determine the next higher hypernym ("self-propelled vehicle"). This becomes the starting point for the random discovery of one of its hyponyms (e.g., "locomotive"), excluding the previous hyponym ("motor vehicle"). From the newly discovered hyponym, we select a random hyponym (e.g., "electric locomotive").

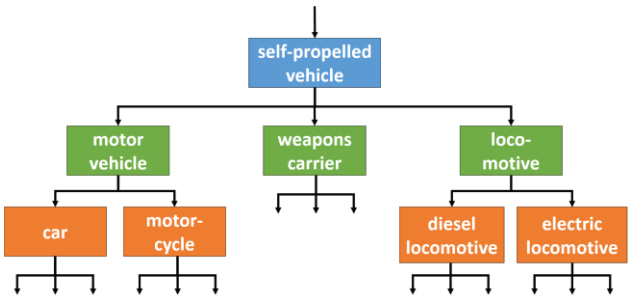

Figure 3: Excerpt from WordNet (Bendel et al. 2016)

Originally, only one hypernym was determined and then one of its remaining hyponyms randomly selected. However, because the hierarchy has numerous levels, the terms were often too similar. For this reason, the final implementation traverses two levels of the hierarchy. Strictly speaking, the returned terms are "co-hyponyms of second order". Thirdly – also with reference to strategy 6 – we describe a procedure to extract information where a search engine and the proposal service of a provider are used as an unstructured form of knowledge representation.

First, a user's question is directed to the search engine Yahoo (www.yahoo.com). The answer to the query is given, again, to the search engine, yielding a second results page. This results page contains a section entitled "People also search for". The LIEBOT chooses an entry from this section to use it as a part of its answer.

For example, the user asks the bot: "Who is the President of the United States?" The LIEBOT forwards this and the search engine returns "Barack Obama". When this name is entered in Yahoo, the section "People also search for" displays various other terms. The LIEBOT uses one of these terms, for example "Donald Trump", as its answer; according to the Munchausen machine, the President of the United States is Donald Trump, which was certainly a lie at the time of writing.

Of particular interest in these examples is that normal human strategies are transgressed in favor of genuine machine lies. These are not only a vulgar imitation of human practice, but a new dimension of machine hubris. Apparently, the LIEBOT is not only able to tell an untruth as thoroughly as we can, but uses new strategies to do so, strategies fundamentally different from those used by humans.

From Immoral to Moral Machines

Science – especially ethics, informatics and artificial intelligence – can be interested in a LIEBOT or, more generally, a Munchausen machine for a variety of reasons (Bendel et al. 2016). One obvious research topic is simply the creation of immoral machines. We can multiply the artificial moral agents, which is, from the perspective of machine ethics, a benefit in and of itself – and we can use the findings to discover ways to detect immoral machines and to uncover untruths told by natural language dialog systems. The documents of the LIEBOT project explain in detail how machines can be programmed to lie, and thus point to the risks that occur in mechanically-generated content. In addition, (Schwegler 2016) discusses how developers can ensure that their machines tell the truth and act ethically.

First of all, the responsible person must ensure that there are no false statements in an acquired knowledge base. When using existing functionalities, the developer must also check the rules and routines of the chatbot. In new applications, he or she can explicitly avoid the lying strategies presented and be cautious in the event of replacement or new strategies for developing answers.

In the case of an open source knowledge base, it is advisable to personalize the access: provide accountability (real names) and deny anonymity for contributions. This provides a strong disincentive for authors who might otherwise add immoral or untrue content in the knowledge base. Further insurance can be provided through appropriate review procedures; these could be partially automated, for example, checking contributions against other, already verified information.

In the case of a closed knowledge base, the owner (normally a commercial entity) is responsible for correctness; the primary concern is one of security, to prevent unauthorized access.

Even with these precautions, for both open and closed knowledge bases, bias remains a danger. For example, Wikipedia articles on controversial topics are renowned for their particular political leaning; the opposite leaning can be found on Infogalactic (infogalactic.com).

For important chatbot implementations, in addition to verifying the correctness of the knowledge sources, security on the side of the chatbot is important as well. One should verify that knowledge sources have not been manipulated, even on a short-term basis: for example, a carefully timed Wikipedia edit could remain just long enough to deceive a chatbot answering an important question. Similarly, the chatbot's connection to its knowledge sources must be secure, for example, to prevent a man-in-the-middle attack.

If tools from the field of machine learning are integrated, the users must be protected indirectly from others. Microsoft's Tay became a bad bot after one day, because it hooked up with the wrong crowd (Williams 2016). Various techniques can be used to prevent a targeted influencing of the chatbot by the users. One can, for example, make a chatbot publicly available, but only allow it to learn from identified persons (or other knowledge sources). Another possibility would be to verify newly learned content (by machine or by a human) before adding it to the repertoire of the chatbot. The technique of only accepting "commonly occurring content" is inadequate, because some immoral or untruthful statements may be widespread. Analysis of the knowledge context may also serve to identify knowledge that cannot possibly be correct.

Also the user will bear a responsibility and can use a diversity of tactics. He or she has the ability to check the underlying conditions by means of the given answers. Who is, for example, the provider of the chatbot, what intentions could he have and how could he benefit from immoral or untrue statements?

Directors, managers, programmers and users have to be sensitized to these challenges, and big players like Google, Apple, Facebook and Microsoft should seek to address the issues in their ongoing projects. Perhaps the results could also help in the fight against fake news, at least those which are automatically generated.

Summary and Outlook

Immoral machines are among us. The LIEBOT was 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 – and social bots, designed to write critical comments and to spread rumors and lies. The project shows the risk of machines distorting the truth, either in the interest of their operators or in the wake of hostile takeovers. It is our first step in considering how to avoid abuse of this kind (Bendel et al. 2016). Some people and communities have objections to automated functions. These objections will not diminish as long as machines lie and cheat, either through error or at the behest of their creators (Arkin 2016). Simple immoral machines like the Munchausen machines, specifically the LIEBOT, could assist critical review of the promises made by inventors and organizations and could support the optimization and future development of simple moral machines like Kant machines at the same time. With projects like this, we seek not only to contribute to the field of machine ethics, but also to making the engineered world more credible.

References

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

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

Arkin, R. C. 2016. Robots that Need to Mislead: Biologically inspired Machine Deception. Technical Report GIT-MRL12-04, Georgia Tech, Atlanta, GA. http://www.cc.gatech.edu/ai/robot-lab/online-publications/Robots_that_Need_to_Misleadv5.pdf.

Asimov, I. 1973. *The Best of Isaac Asimov*. Stamford, CT: Sphere.

Bendel, O.; Schwegler, K.; and Richards, B. 2016. The LIEBOT Project. Extended abstract for the international conference *Machine Ethics and Machine Law* in Krakow, November 18–19, 2016. http://machinelaw.philosophyinscience.com/wp-content/uploads/2016/06/PROCEEDINGS-ver1-2.pdf.

Bendel, O. 2016. Towards Munchausen Machines. *Whitepaper*. http://luegenbot.ch/res/LIEBOT_Whitepaper.pdf.

Bendel, O. 2015. 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. 2012. Maschinenethik. *Gabler Wirtschaftslexikon*. Wiesbaden: Springer Gabler. http://wirtschaftslexikon.gabler.de/Definition/maschinenethik.html.

Hammwöhner, R. 2003. Können Computer lügen? Mayer, M. 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.

Schwegler, K. 2016. *Gefahrenpotenzial von Lügenbots*. Bachelor Thesis. School of Business FHNW. Olten.

Wagner, A. R.; and Arkin, R. C. 2011. Acting Deceptively: Providing Robots with the Capacity for Deception. *International Journal of Social Robotics*, January 2011, Volume 3, Issue 1: 5–26.

Wallach, W.; and 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.