Robots between the Devil and the Deep Blue Sea

Robôs: entre o diabo e o profundo mar azul

Oliver Bendel*

RESUMO

Este artigo apresenta dilemas clássicos e os transfere para a era da informação, focando particularmente o uso problemático de chatbots, robôs, drones e veículos autodirigidos. Soluções de conceitos são desenvolvidas na perspectiva de ética das máuinas, entre outros. Vemos que os dilemas clássicos são úteis para enfrentar os desafios atuais e ajudam a discutir as opções decisórias de sistemas parcial ou totalmente autônomos e para sintetizar a robótica, a inteligência artificial e a ciência da computação nessas questões para otimizar seus resultados e produtos.

Palavras-chave: Robótica; Robôs de serviço; Veículos autodirigidos; Chatbots; Ética das máquinas.

ABSTRACT

This article presents classic dilemmas and transfers them to the information age with special focus on the problem-ridden use of chatbots, robots, drones and self-driving cars. Solution concepts are developed from the perspective of machine ethics among others. It turns out that classic dilemmas are useful for mastering today's challenges and helpful for discussing the decision-making options of partly or fully autonomous systems and for sensitizing robotics, artificial intelligence and computer science to such matters in order to optimize their results and products.

Keywords: Dilemmas; Robotics; Service Robots; Self-driving Cars; Chatbots; Machine Ethics.

INTRODUCTION

The dilemma of Euathlos and his famous mentor Protagoras, whom he was meant to pay with the proceeds of his first victory in court ("The Paradox of the Court"), has long since held a permanent position in the history of philosophy. None such victory happened and Protagoras sued Euathlos. So who gets the money or keeps it and for what reason? The Heinz-Dilemma by Lawrence Kohlberg, author of the theory of the stages of moral development, is known to a wider public. Would stealing a grossly overpriced drug to rescue one's terminally ill partner be permissible if doing so damages the inventor of the drug who aims at maximum profit? The Prisoner's Dilemma, Buridan's Ass, and the Plank of Carneadas – a lot of terms, examples and images have been coined between the classic and the modern age.

In fact constructing and varying dilemmas is a favorite pastime of real and wannabe philosophers. Other than logic, ethics particular loves doing so. As a discipline with morality as its object it deals with decision-making in the context of good or bad

* Doutor em Sistemas de Informação, pela Universidade de St. Gallen, na Suíça. University of Applied Sciences and Arts , Northwestern Switzerland FHNW, School of Business, Institute for Information Systems, Bahnhofstrasse 6. CH-5210 Windisch. Telefone: +41 56 202 73 16.

intentions, good or bad lives, and justice and injustice. Dilemmas are plights or predicaments where choosing between two options seems to be very difficult, or per se leads to unwanted results. People who get to know and rethink such thought experiments often find them paradox and eccentric. Depending on the perspective, some of the thought experiments can be interpreted as dilemmas or trilemmas.

Chatbots, robots, drones and self-driving cars are the starting point for this article. They get into classic dilemmas which are re-interpreted here. Solution concepts developed from the perspective of machine ethics are discussed, and questions from the perspective of information ethics and technology ethics are raised. Robotics and artificial intelligence (AI), which have to work on the technical implementation and on relevant fields of application, might benefit, as well as computer science and business administration. Ethics and robotics will be investigated in more detail in the next two chapters.

SPECIFIC FIELDS OF APPLIED ETHICS AND MACHINE ETHICS

Ethics is a discipline of philosophy, and morality, or in other words the normative framework of behavior towards other human beings, towards one's self and towards the environment, is its object. Aristotle is considered the founder of this discipline in the classic age, while Immanuel Kant is considered its renewer in the enlightenment period. He focused on the rationality and autonomy as core issues. Arthur Schopenhauer and Jeremy Bentham referred to the ability of empathy and suffering, and therefore integrated animals as objects of morality. Ethics as science for instance uses logical, discursive, dialectical or analogous methods (PIEPER, 2007).

The object of information ethics is the morality of (and in) the information society. It analyses how its members when offering and using information and communication technologies or digital media behave or should and want to behave (BENDEL, 2012b). Technology ethics refers to moral issues of the use of technique and technology. It might deal with vehicles or weapons as well as with nanotechnology. The transitions to information ethics today are seamless.

The aforementioned specific fields of applied ethics are part of human ethics (humans are the subjects of morality) and they form applied ethics which analyses problems and solutions in practice. The object of machine ethics is the morality of machines (ANDERSON; ANDERSON, 2011), and in particular the morality of (partly) autonomous systems such as agents, certain robots, certain drones, and self-driving cars (BENDEL, 2012a). It can be considered part of information and technology ethics or an equivalent to human ethics.

In the end, applied ethics and technology assessment should ask fundamental questions (BENDEL, 2014, 326): How should (information) technology be in the future? Do we want to have (partly) autonomous systems at all? Do we want to have machines that attain a consciousness and think and feel? That behave morally, as subjects of morality, and that are even objects of morality (with intrinsic values and rights) some day? Hans Jonas' precautionary principle seems to be as modern as never before (JONAS, 1979). He calls for the avoidance of inassessable risks, so as not to undermine the continuity of mankind. Perhaps autonomous systems are risks of this nature and must be banned.

TRANSFERRING CLASSIC DILEMMAS

Robotics, or robot technology, has come up in the 1950s. It deals with the draft, design, control, production and operation of robots. In a wider sense this term includes not only industrial robots and service robots but also certain drones and self-driving cars. With humanoid robots the production of extremities and skin, facial expressions and gestures as well as natural language skills might be an issue. A major problem of this form of robotics is the "uncanny valley". The artificial has to be very close to the natural in order not to irritate the beholder. Pure software robots, so-called agents or bots, have certain advantages in this respect. They are developed for instance by AI and computer science. Robots and bots are part of the machinery being the object of machine ethics.

Dilemmas come up frequently in the fields of application of robotics, AI and information technology. Often the trolley problem is applied to traffic with self-driving cars (BENDEL, 2015c). Below four classic dilemmas are presented and transferred into the information age, wherein trolley problem and fat man problem may be considered variants. Concepts for solutions are sketched and discussed without going into details of their technical realization. There is no doubt that machines cannot always master predicaments with satisfactory results, often they will be found to have more problems than humans.

The robot car problem

The trolley problem is a thought experiment conceived by the British philosopher Philippa Foot.[1] A railroad trolley gets out of control and speeds toward five persons. It could be diverted to another track on which one other human being is standing. Can one tolerate the death of this one person in order to save the lives of the people in the group? This question resurfaces in the traffic of the future (BENDEL, 2015b). The brakes of a self-driving car fail. The alternatives might be similar to those in the trolley problem. The car speeds towards a group of five persons. It could bypass the group to one side, where one man is standing, obviously unable to leap away. If the car drives on straight, casualties will be unavoidable. How to decide in this plight? The robot car problem is essentially identical to the original problem.

The fat man problem offered by US-American philosopher Judith Jarvis Thomson is a variant of that problem. The fat man who gave the problem its name is pushed onto the tracks in order to stop the trolley before it hurts the group of five persons. This means the death of a person is not only tolerated, it is caused intentionally. This problem too can be transferred to the self-driving car with failing brakes (BENDEL, 2015b). It is further assumed that the steering is malfunctional and permits only minimum deviations. Two groups are on the lane; in front of one group is a fat man who would stop the car. May the car kill this one person to save the group? Or does it have to steer towards the other group with the risk of causing more casualties? In this version of the robot car problem, thoughts are related rather than identical.

The literature on machine ethics gives several indications towards possible solutions (ANDERSON; ANDERSON, 2011). In general, the conflicts can be tackled following defined rules, moving within a duty ethics. Reflections on the immediate future might

[1] The thought experiment was presented in the 1960s and is widely and intensively acclaimed until today (Foot, 1967).

come along. Cases are compared and consequences are analyzed, with the ethics of consequences being the normative framework. In a student paper mentored by the author a formula was developed that benchmarks the alternatives and determines what seems to be the best possible solution (BENDEL, 2014, 324). The criteria for benchmarking are age, health, number of persons etc. Causing the death of as few persons as possible follows classic utilitarian principles. Applying these principles, it might be permissible under certain circumstances to actively kill one person in order to save a group. The pertinent literature and certain companies also propose random generators (BEUTH, 2014). The overall principle is "One against all". A person or a car is causing an accident and has to select the suitable victim. The option of self-destruction occurs only at the margins of the discussions. Who causes the accident bears the responsibility – literally or figuratively. Uninvolved persons are not killed, but the car destroys itself and the passengers on board. In this case, one might again consider how many victims this solution would cause in consequence.

Technology and information ethics do not offer ready-to-use solutions but provide insight into the role of technology, for instance in relation to the autonomy gain of machines and autonomy loss of humans or the handling of information before, during and after an accident. The question is: are we to succumb to technology and are we only the sum of our data in certain situations? Is it permissible to manipulate this data in order to be the victor of an accident? This question has to be discussed along with error probability and proneness to erring of moral machines.

Carneades' lay-by

The thought experiment about Carneades' plank probably dates back to the philosopher from Cyrene.[2] A plank is floating next to two shipwrecked persons, but it can bear just one of them, not both. One of them kills the other and survives. The question is: can this murder be justified, and if yes, how so? Similar constellations could occur in modern traffic (BENDEL, 2015b). A ghost driver steers a heavy truck along the freeway. A self-driving car can save its driver only by reaching a tiny lay-by. Another self-driving car is claiming the same lay-by at the same time, also trying to save its owner. May one of them kill or let kill the "adversary" to make sure its owner survives? This transfer reminds us of the classic dilemma in several aspects.

How could the machine make a reasonable decision? Again, formulae might be an option. Age, health and number of passengers could be useful criteria for evaluation. Saving as many passengers as possible would follow classic utilitarian principles, while modern utilitarianists would ask for the interests of the involved and non-involved players. "Cars casting the dice" might be another option. All in all, such options seem to be less cynical than with the self-driving car problem where – as already mentioned – the principle is "One against all". Carneades' lay-by is a duel for naked survival, and the situation might even classify as self-defense situation. The principle is "He/She or I".

Technology ethics and information ethics again offer some orientation on the role of technology, for instance in relation to the benchmarking for decision-making of different autonomous systems and the opportunity to buy into decisions as the owner of the car. Once more it has to be discussed whether or not it is permissible to manipulate data in order to be the victor of the conflict, while the error probability

[2] The dilemma is mentioned in Cicero's "De officiis", 3.89–90.

(and proneness to erring) of moral machines is another discussion-worthy item. The term of digital self-defense acquires new meanings, such as the concept of informational self-defense, which normally only covers self-protection and use of force without lethal consequences.

Buridan's robot

Buridan's ass starves between two bundles of hay because it cannot make up its mind on which one to eat.[3] The core of this dilemma can be traced back to Aristotle. Variants were created by the Persian philosopher Al-Ghazālī and by Johannes Buridan or his critics who added the ass as a metaphor to the thought image. The question is: how can a machine remain capable of making decisions when similar stimuli affect it (BENDEL, 2013a)? A service robot in a museum or in a store is addressed by two or three customers at the same time. Whom shall it serve first? The richest one or the busiest one? Or those who are older, more important or louder? Or should the decision in such a dilemma or trilemma be made at random? A combat robot tracks down a terrorist he is programmed to eliminate. Suddenly the terrorist's twin appears who is virtuous in thoughts and deeds, and identical in looks. Whom should the machine kill, one of the brothers, both or none? There is the risk of the good one dying and the bad one getting away. On second glance, Buridan's robot is quite similar to Buridan's ass.

How to prevent that the robot dies "from hunger or thirst" between the twins and gets incapable of making a decision? While at the same time keeping up its morally correct behavior? One could teach the combat robot not to trust its first impression without reservations (BENDEL, 2013a). When the stimuli all are equally strong more stimuli are required. This means the robot has to find out more information about its target objects. It might ask the brothers Solomonic questions, it could take their fingerprints and compare them to available data. These are elaborate but promising strategies. The twin brothers definitely have to be kept prisoners until the decision is made. Of course this might overtax the robot, for instance if the robot is a UAV without arms and hands. The specified rule obviously would have to be modified. The description must not only be accurate, it also has to match the searched person.

Technology ethics and information ethics are required to define the framework in civil and military settings. If customers seek preferred treatment they will probably try to influence the robot by making misrepresentations and giving false answers to its questions. Interpreting human cunning correctly and responding to it in a way that protects human dignity will not be easy for robots. When correct identification and elimination of target persons is at stake, manipulation of their data has to be expected. The question is what does this mean to one's own identity and to the integrity of others? Military ethics has to find an answer to the question how just a war, or fighting against terrorists, with autonomous systems is.

The liarbot problem

An old dilemma has been discussed in so-called Holy Scriptures as well as in the works of philosophers from Socrates to Kant: lies are banned but white lies are commonly

[3] Aristotle described the paradox in "On the Heavens".

tolerated as exceptions in certain situations. John Stuart Mill considers the love of truth useful and weakening it detrimental. He says one has to evaluate carefully according to the principle of utility from case to case (MILL, 1976, 39 f.). According to Kant being honest in all declarations is a rule of reason not to be limited (KANT, 1914, 429). Like most other machines, chatbots will usually tell the truth, not for moral but for pragmatic reasons. They provide services meant to entertain, support and inform people. They would not function properly if they were not reliable in terms of truth. The liarbot (BENDEL, 2013b), which knows the truth but construes an untruth, is a counterdraft, even if it construes white lies only.

The idea of a GOODBOT, developed as a prototype from 2013 to 2014 at the School of Business FHNW by students of the author (AEGERTER, 2014), was born together with seven meta rules. Number 3 was: "The GOODBOT shall not hurt the user through its appearance, facial expressions and movements, or words." And number 4: "The GOODBOT does not lie to the user, or it makes clear to the user that it lies." The students were allowed to adjust the author's catalog. Their solution was that the GOODBOT generally tells the truth with one single exception: If rule 3 is disobeyed, the machine first will try to by-pass. If the counterpart insists, the GOODBOT will lie to it (BENDEL, 2015a). The Bot sticks to the truth, quite strictly so, but just like Mill it knows exceptions. The overall goal of the GOODBOT is morally adequate behavior towards its user. It is able to escalate over several steps. In extreme cases it will give out a national hotline number for the country where the human chat partner resides.

The aforementioned rules relate to the morality of machines, and they describe the assumed behavior and the intended well-being of the user. In so far it seems reasonable to involve technology ethics and information ethics. The question is: shall the GOODBOT encourage the chat partner to avoid certain topics? The user might be instructed by the machine and limited in his respectively her autonomy. The analysis of the IP address which is required for correct disclosure of the national emergency hotline might infringe against informational self-determination while the benefit it provides is big, this creates a new dilemma.

CONCLUSION AND OUTLOOK

More dilemmas wait for being transferred into the 21st century, to the world of software agents and service robots, unmanned combat aerial vehicles and self-driving cars. They can be taken from the works of classic and modern philosophers or from the works of science fiction writers such as Isaac Asimov and Stanisław Lem. They are meant for people who want to apply their minds to reviewing their actions. Of course we are not restricted to references to traditions and fictions of the great masters. We can make new thought experiments of our own. Not only we humans can – machines can do so too. Thinking about dilemmas where they are the protagonists might make them more humble, leaving decision-making to humans when in doubt (BENDEL, 2015b).

Machine ethics can present possible solutions for moral machines to enable them to decide in dilemmas. Maybe their decisions will not always be correct, but they will be bearable and understandable (WALLACH; ALLEN, 2009). Robotics and artificial intelligence could both benefit. The machines they produce face classic problems, overcome serious challenges of our modern age, and might gain acceptance in the discerning information society. In order to achieve this goal technology ethics and information ethics have to be involved. They will steer the perspective away from the machine towards the human being, and ask for correct human behavior and a good

human life. And how human beings feel in times when machines seem to take over decision-making. Not only numerous new subjects of morality come into the world, the morality of human beings is measured against the morality of machines more often. This can lead to rationalization but also to its opposite, in as far as the machine becomes the competitor against whom to defend and demarcate oneself. Without doubt this situation will lead to new dilemmas.

Last but not least we have to ask whether duty ethics and ethics of consequences are preferable models of normative ethics in this context or whether other models are possible. According to Pieper (PIEPER, 2007, 270), seven fundamental models can be distinguished, the transcendental, the existential and the eudemonistic approach, the contracting theory, and the traditional, the materialistic and the life-world model. In the context of machine ethics, the traditional model – which seem to be a promising candidate in this context, considering the GOODBOT – may mean that a machine needs to acquire virtues such as wisdom, justice, courage and temperance, and develops a character which includes a set of them. The "morally right" action implicitly follows from this character, i.e, from the interaction of virtues. Similar to the rule-based approach, its virtues may be prioritized or formed in a special way in order to adjust them to the intended character of the machine. However, a human character also includes assertiveness, empathy, and intuition. It is a reasonable assumption that it is difficult to create something beyond a "virtue machine". And after all, it remains unclear if virtuousness could help solve the robot car problem or Carneades' lay-by.

LITERATURE

AEGERTER, J. FHNW forscht an "moralisch gutem" Chatbot. In: Netzwoche, 4/2014, p. 18.

ANDERSON, M; ANDERSON, S.L. (eds.). Machine Ethics. Cambridge: Cambridge University Press, 2011.

BENDEL, O. Können Maschinen lügen? Die Wahrheit über Münchhausen-Maschinen. In: Telepolis, March 1, 2015a. Disponível em: <http://www.heise.de/tp/artikel/44/44242/1.html. Acesso em: 31 mar. 2015.

BENDEL, O. Die Parkbucht des Karneades: Viereinhalb Dilemmata der Robotik. In: inside-it.ch, March 17, 2015b. Disponível em: <http://www.inside-it.ch/articles/39531>. Acesso em: 31 mar. 2015.

BENDEL, O. Selbstständig fahrende Autos. Beitrag für das Gabler Wirtschaftslexikon. Gabler/Springer, Wiesbaden 2015. Disponível em: <http://wirtschaftslexikon.gabler.de/Definition/selbststaendig-fahrende-autos.html>. Acesso em: 31 mar. 2015c.

BENDEL, O. Towards Machine Ethics. In: MICHALEK, T.; HEBÁKOVÁ, L.; HENNEN, L. et al. (eds.). Technology Assessment and Policy Areas of Great Transitions. 1st PACITA Project Conference, March 13 – 15, 2013c. Prague 2014, pp. 321 – 326.

BENDEL, O. Buridans Robot: Überlegungen zu maschinellen Dilemmata. In: Telepolis, November 20, 2013a. Disponível em: <http://www.heise.de/tp/artikel/40/40328/1.html>. Acesso em: 31 mar. 2015.

BENDEL, O. Der Lügenbot und andere Münchhausen-Maschinen. In: CyberPress, September 11, 2013b. Disponível em: <http://cyberpress.de/wiki/Maschinenethik>. Acesso em: 31 mar. 2015.

BENDEL, O. Maschinenethik. Gabler Wirtschaftslexikon. Wiesbaden: Gabler/Springer, 2012a. Disponível em: <http://wirtschaftslexikon.gabler.de/Definition/maschinenethik.html>. Acesso em: 31 mar. 2015.

BENDEL, O. Informationsethik. Gabler Wirtschaftslexikon. Wiesbaden: Gabler/Springer, 2012b. Disponível em: <http://wirtschaftslexikon.gabler.de/Definition/informationsethik.html>.

BEUTH, P. Wenn Software über Leben und Tod entscheidet, In: ZEIT ONLINE, Mai 13, 2014. Disponível em: <http://www.zeit.de/digital/internet/2014-05/unfall-fahrerlose-autos-ethik>. Acesso em: 31 mar. 2015.

FOOT, P. The Problem of Abortion and the Doctrine of the Double Effect. In: Oxford Review, Number 5, 1967.

HÖFFE, O. Lexikon der Ethik. 7nd edition. München: C. H. Beck, 2008.

JONAS, H. Das Prinzip Verantwortung: Versuch einer Ethik für die technologische Zivilisation. Frankfurt am Main: Suhrkamp, 1979.

KANT, I. Werke (Akademie-Ausgabe), Bd. 6. Berlin: Königlich Preußische Akademie der Wissenschaften, 1914.

KUHLEN, R. Informationsethik: Umgang mit Wissen und Informationen in elektronischen Räumen. Konstanz: UVK, 2004.

MILL, J.S. Der Utilitarismus. Ditzingen: Reclam, 1976.

PIEPER, A. Einführung in die Ethik. 6nd edition. Tübingen und Basel: A. Francke, 2007.

WALLACH, W.; ALLEN, C. Moral Machines: Teaching Robots Right from Wrong. Oxford: Oxford University Press, 2009.