Machine Ethics and Machine Law

Jagiellonian University

November 18, 2016 – November 19, 2016

Cracow, Poland

E-Proceedings

Organisers:

Jagiellonian University (the Institute of Philosophy and the Department for the Philosophy of Law and Legal Ethics),

Copernicus Center for Interdisciplinary Studies

Jagiellonian University’s Cognitive Science Student Association

Copernicus
Center

Organising Committee:

Bipin Indurkhya (co-chair) Jagiellonian University;

Bartosz Brożek (co-chair) Jagiellonian University;

Georgi Stojanov (co-chair) American University in Paris

Scientific Program Committee

Peter Asaro (New School of New York)

Joanna Bryson (University of Bath, UK)

Elizabeth Kinne (American University in Paris)

Tom Lenaerts (Université Libre de Bruxelles and Vrije Universiteit Brussel)

Cindy Mason (Stanford University)

Luis Moniz Pereira (Universidade Nova de Lisboa, Portugal)

Henry Prakken (Utrecht University and University of Groningen)

Kathleen Richardson (De Montfort University, UK)

Antonino Rotolo (University of Bologna)

Tony Veale (University College Dublin)

Local Organising Committee:

Joanna Ganczarek, Marek Jakubiec, Bartosz Janik, Michał Klincewicz, Bartłomiej Kucharzyk, Łukasz Kurek, Kamil Mamak, Katarzyna Marchewka, Rafał Michalczak, Tomasz Zygmunt, Tomasz Żuradzki

Table of contents

20. Justyna Zygmunt. To teach a machine a sense of art – problems with automated methods of fighting copyright infringements on the example of YouTube Content ID.

AI and states of mind from a legal perspective: from intentional states to guilt

José Maia Neves

Informatics Department
School of Engineering
University of Minho
Campus de Gualtar,
Braga, Portugal

Francisco Andrade

Law School
University of Minho
Campus de Gualtar,
Braga, Portugal

Braga, Portugal Miguel Freitas

Law School
University of Minho
Campus de Gualtar,
Braga, Portugal

Abstract

The autonomy of software agents and the consideration of its intentional states have led us to consider the issue of divergences and defects in the declaration of will of the software agent. Nevertheless, there is a topic that still requires a legal and artificial intelligence combined analysis since the acting of autonomous software agents brings along the issues of guilt and negligence. In this paper we will try to identify the concepts of guilt and negligence and its various different levels, both in civil and criminal domains, and from these we will try to enquire if there is any possibility of developing a system of knowledge representation and reasoning, under a formal framework based on Logic Programming, allowing the evaluation and representation of the possible levels of guilt in the actions of autonomous software agents.

In order to accomplish this representation, we have to take a look at the civil and criminal general legal theory on guilt and to distinguish the different possible levels of guilt encompassed in the actions of software. Of course we are aware that often it will not be an easy task to distinguish different behaviours and to relate them with different environmental conditions and, from these, to try to understand the reasons that led the software to adopt a certain behaviour (Sartor, 2009; Pereira & Saptawijaya, 2016)[1] – it is understandable that there may be different conditions in similar situations, with intricate relations and the available data may be incomplete, contradictory and/or default, either in qualitative or quantitative terms. In order to overcome these drawbacks, we shall have to apply reasoning techniques and knowledge representation to set the structure of the information and the associate inference mechanisms. We will use a Logic Programming based approach to knowledge representation and reasoning, complemented with a computational framework based on Artificial Neural Networks (Martins *et al*, 2015).

Nowadays, software is no more just an instrument that humans use in order to accelerate the speed at which electronic transactions occur. In the new platforms of electronic commerce, software has a much more proactive role, since it can initiate negotiation and play different roles related to negotiation and entering into contracts, often without any human intervention. Software became thus an active participant in commerce (Weitzenboeck, 2001) and humans may not have the consciousness that a contractual negotiation was initiat-

[1] The complexity of recognizing intentions in software behavior must be acknowledged (Pereira & Saptawijaya, 2016:143-144).

ed or the contract celebrated[2]. This can lead to divergences and defects in the declaration of will. From a legal standpoint, article 483 of the Portuguese Civil Code usually requires "*dolus* or mere guilt" for the consideration of liability, being the possibility of liability without guilt exceptional and only considered whenever expressly accepted by law. Traditionally, guilt has been defined as a criteria of imputation of the act to the agent (Leitão, 2013). A requirement for this imputation would be either (in a psychological sense) the consideration that the act arises from the (free) will of the agent, or (in a normative sense) as judgement of the actor being blamed by his behaviour (Leitão, 2013). We must also consider the moral permissibility of certain actions mainly in dilemma situations (Pereira & Saptawijaya, 2016) For this connection to be established it must be stated that the agent adopted a certain behaviour but, according to law, should have adopted a different behaviour (Leitão, 2013), although in the case of software agents it might be questionable whether or not software agents are supposed to abide to legal norms[3].

In this normative sense we should have to consider a concept of "due diligence" (Leitão, 2013) which may be difficult to encompass in the case of the acts of software agents. Greater difficulty arises when discussing the possibility of criminal liability of software agents. The concept of guilt in criminal theory is one riddled with doubts that fuelled great debates among legal theorists. In common law, criminal liability depends upon proving not only the *actus reus* but also the *mens rea* of a crime. Each crime can be divided into two elements, *actus reus* and *mens rea*, the first being translated into a guilty act and the second into a guilty mind. So, there must be a blameworthy state of mind if someone is to be held criminally liability. This state of mind can either be intention (direct and oblique), knowledge, recklessness or negligence (Ormerod & Laird, 2015), depending on how the crime is worded in law and its legal requirements. When we shift our attention towards civil law countries, the blameworthy states of mind are categorized differently: *dolus*, which is divided into *dolus directus*, *dolus indirectus* and *dolus eventualis*, and negligence, split into conscious (advertent) negligence and unconscious (inadvertent) negligence (Dias, 2012). The frontier between different states of mind is frequently hard to trace, especially for the civil law judge when in doubt between *dolus eventualis* and conscious negligence. For this reason we propose a Logic Programming based approach to knowledge representation and reasoning, complemented with a computational framework based on Artificial Neural Networks, as a tool for judicial decisions.

References

Dias, Figueiredo. 2012. Direito penal, Parte Geral, Tomo I, Questões fundamentais, A doutrina geral do crime. Coimbra: Coimbra Editora.

Martins, Rosário, Mendes, Teresa, Grañeda, José, Gusmão, Rodrigo, Vicente, Henrique, and Neves, José. 2015. Artificial Neural Networks, in Acute Coronary Syndrome Screening, in Bioinformatics and Biomedical Engineering, Springer, pp. 108-119.

Leitão, Luís Menezes. 2013. Direito das Obrigações. Coimbra: Almedina.

Ormerod, David, and Laird, Karl. 2015. Smith and Hogan's criminal law (14. ed.). Oxford: Oxford University Press.

Pereira, Luís Moniz, and Saptawijaya, Ari. 2016. Programming Machine Ethics. Springer.

Sartor, Giovanni. 2009. Cognitive Automata and the Law: Electronic Contracting and the Intentionality of Software Agents. Artificial Intelligence and Law, 17, 4, 253-290

Weitzenboeck, Emily. 2001. Electronic agents and the formation of contracts, in International Journal of Law and Information Technology, Vol. 9, No. 3, pp. 204-234.

[2] The Portuguese legal framework on the services of information society, particularly electronic commerce (DL 7/2004), already recognizes in article 33 that electronic contracting may occur without human intervention.

[3] See the important "trolley problem".

The LIEBOT Project

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

This paper describes the foundations of the LIEBOT project, whose objectives are to implement a lying chatbot and to investigate the chances and risks of immoral machines.

Introduction

The category of simple immoral machines includes so-called Munchausen machines (Bendel 2015), that is to say machines and systems that systematically produce lies. A concrete manifestation of this category is a chatbot that tells an untruth, like the LIEBOT. The LIEBOT project, which is discussed in this paper, is based on preparatory works by the scientist who already initiated the GOODBOT, a simple moral machine (Bendel 2013a). A business informatics student was contracted in early 2016 to implement the LIEBOT as a prototype in the scope of his graduation thesis, as an extension of the preparatory works.

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 (Aegerter 2014). Can these systems be trusted? Do they always tell the truth? It is possible for producers and providers to avoid Munchausen machines and for users to detect them.

The LIEBOT is available as a chatbot on the website liebot.org. It has been programmed in Java, with the Eclipse Scout Neon Framework. The chatbot tells lies in areas of all kinds, and concentrates on two specific fields of application: energy drinks and Basel as a tourism region. It has a robot-like, animated avatar whose nose for example grows like Pinocchio's if an untruth is produced.

Lying Machines

Whether or not machines are really capable of lying to us (or to other machines) is the subject of controversial discussion. 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 famous 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”. Further research topics are automatic deception and misleading (Wagner and Arkin 2011; Shim and Arkin 2013) and machines that suggest statements which may be true and false and that learn by human feedback like the Twitter bot Nell (user name @cmunell).

Strategies of Lying

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 or assuming the truth, it constructs an untruth.

In (Schwegler 2016) a total of ten strategies are described:

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 appear, but does not have to. The majority of the strategies were implemented in the LIEBOT project, sometimes in combination. They were also equipped with different probabilities, so that lying does not always occur in the dialogs. To illustrate the implementation, we explain strategy 6 partly in detail: the exchange of terms with antonyms and co-hyponyms, as well as methods of information extraction.

First, we describe the implementation of the production and use of co-hyponyms, based on WordNet (Princeton University). 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 generation of co-hyponyms, but carries this one step farther: rather than generating 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. 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").

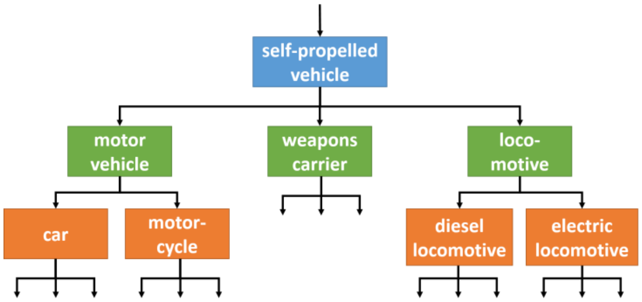

Fig. 3: Excerpt from WordNet

Originally, only one hypernym was determined and then one of its remaining hyponyms randomly selected. However, because the hierarchy has many levels, the terms were often too similar. For this reason, an implementation over two levels of the hierarchy was carried out. Strictly speaking, the returned terms are co-hyponyms of second order.

In the following – 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. The answer to the query is given, again, to the search engine. The result page of the second search request contains a section entitled "People also search for". The LIEBOT chooses the first entry from this section for further processing. For example, the user asks the chatbot: "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 in summer 2016.

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.

From Immoral to Moral Machines

Science can be interested in a Munchausen machine for a variety of reasons. One obvious research topic is simply the creation of immoral machines. We can multiply the moral agents, which is, from the perspective of machine ethics, a benefit for itself – and we can use the findings to discover ways to detect bad machines, and to uncover untruths told by natural language dialog systems.

The LIEBOT project explains in detail how machines can be programmed to lie, and thus points to the risks that occur in mechanically-generated content. In addition, (Schwegler 2016) discusses how developers can ensure that their machines tell the truth by accessing reliable sources and protecting knowledge bases from enemy attacks, as well as how users can recognize such systematically lying machines.

Conclusion and Outlook

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

This research is our first step in considering 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. Simple immoral machines like the Munchausen machines, specifically the LIEBOT, could assist critical review of the promises made by persons and organizations and could support the optimization and future development of simple moral machines at the same time. 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.

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

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.

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

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

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

Shim, J., and Arkin, R. C. 2013. A Taxonomy of Robot Deception and its Benefits in HRI. *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*, 2328–2335.

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.

Is Morality the Last Frontier for Machines?

Bipin Indurkhya
Jagiellonian University, Cracow (Poland)
bipin.indurkhya@uj.edu.pl

Introduction: As autonomous decision-making systems are becoming more and more commonplace — from self-driving cars, to military robots, including drones, to social companion robots including sex robots — they are rapidly encroaching into domains where moral and ethical values play a key role. For example, a self-driven car, on sensing a mechanical failure, may have to decide whether to hit some pedestrians, or drive into a ditch thereby risking the lives of its occupants. A military robot may have to decide whether to fire a shell at a house where a terrorist and also five other possibly innocent people are hiding. (This was the theme of a recent movie *Eye in the Sky.)* A companion robot may have to decide whether to lie to the companion human with a terminal disease about whether they will recover. These issues have ignited an intense interdisciplinary discussion on how machines should be designed to handle such decisions, and if machines should handle such decisions at all (Arkin, Ulam & Wagner 2012; Levy 2007; Lin, Abney & Bekey 2011).

Indeed, some researchers have argued that that the domain for ethical decisions is essentially a human forte, and a machine ought not to venture in there (Bryson 2016). They have argued that machines should be deliberately designed to make it obvious that they are machines, so that no moral agency is attributed to them. Not completely sidestepping this debate, I would like to raise two issues that raise doubts about this position.

Emergence of human-robot blends: The current debate on whether machine can be moral agent or not is based on assuming a clear separation between robots and humans: robots are machines designed by humans using mechanical and electronic components; humans are biological beings who are born with some genetic dispositions inherited from the parents, and develop their cognitive functionalities over time. However, this boundary is being blurred slowly. On one hand, people are incorporating robotic components in their bodies and brains to increase their physical and cognitive abilities (Schwartzman 2011; Warwick 2003, 2014). On the other hand, researchers are designing machines and robots using biological material (Ben-Ary & Ben-Ary 2016; Warwick 2010). At the moment, the state-of-the-art is still far from generating a human-robot blend that would be hard to classify as a robot or a human, but it might soon become a reality. In such situations, it would be hard to say who is a moral agent and who is not.

Machines are susceptible for hacking: It is sometimes argued that robots, especially military robots, should not be completely autonomous because they can be hacked (Lin 2011). This, however, is a problem with humans as well. Ever since the dawn of history, there have been many examples where some human was bribed or blackmailed into turning against their own side, or change their moral stance. (Consider Judas, Brutus, Alfred Redl, Harold Cole, Mir Jafar, Aldrich Ames, and so on.) So this cannot be a basis for denying machines the moral agency.

In the rest of this paper, I would like to focus on a more pragmatic issue. Assuming that the development of technology cannot be stopped by making such laws etc. — indeed, there are already companion robots that interact with people in a human-like way and try to fulfill their social needs — the issue I will address is how to make their decisions acceptable to humans. In this regard, I will consider two factors.

***Sophie's choice* effect:** What does a human do when confronted with two choices that are both horrifying? I refer to this as *Sophie's Choice* effect based on the film with this title, where a Nazi officer forces a Polish mother (played by Meryl Streep) to choose one of her two children, whose life would be spared. One can find many similar real-life cases, especially during natural disasters like earthquakes and floods, or during man-made disasters like wars. No matter what one chooses, such decisions usually leave a deep psychological scar and can traumatize the person for the rest of her or his life.

This issue has been explored extensively in recent years as what is known as the trolley problem and its variants (Bruers & Braeckman 2014; Navarette *et al.* 2012; Nucci 2013). These experiments, however, do not reveal what a person would actually do in such a situation, what psychological trauma they will face as a result of it, and how they justify their choices. Sometimes they provide some justification, but then varying the experimental conditions show that they do not necessarily act according to their own justification (Bauman *et al.* 2014). Nonetheless, these experiments provide fodder for how autonomous machines like self-driving car might be programmed with moral rules (Brogan 2016).

Such dilemmas are also faced by governments and social groups, often in war campaigns, in starting big construction projects like dams, in social projects like relocating slums, and so on. In such situations, some public justification is often provided, though, in almost all such cases it is not accepted by everyone. Perhaps the most well known case may be the justification put forth by the US Government for dropping nuclear bombs on Hiroshima and Nagasaki, namely that it saved lives of American as well as Japanese people (Morton 1960). We can learn from such explanations as to what is acceptable or not acceptable by social groups. A very useful case in point is the development of Triage system to determine the priority of medical treatment of patients, which is widely used (Iserson & Moskop 2007; Moskop & Iserson 2007; Robertson-Steel 2006).

Is morality the last frontier? Humans generally show a reluctance to accept machine superiority. Almost always, humans have challenged the machine when it makes a foray into some human domain. There is the legend of John Henry, who competed with a steam-powered hammer, won the contest, but then died immediately after as his heart gave out. Deep Blue defeated the reigning world champion Garry Kasparov in 1997, but some claim that the computer does not 'understand' chess because it does not play as humans do (Linhares 2014). More recently, in 2011, the computer system Watson won the quiz game Jeopardy against the best humans, but many scholars deny that it has any 'understanding' of the questions or related concepts (Searle 2011). So it is not surprising that morality, where machines have yet to demonstrate their superiority in some way, is considered off limit for machines.

When we extract the cognitive mechanism underlying some human behavior, and make an algorithmic version of it, people generally do not accept it. There are several examples that illustrate this human trait. Consider the use of actuarial tables in making parole decisions. Evidence has been put forth to show that actuarial tables are more reliable than human experts, but their role in legal decision-making is still being disputed (Dawes, Faust & Meehl 1989; Krauss & Sales 2001; Litwack 2001; Starr 2014).

There are two major limitations of these statistical methods, or algorithmic methods based on behavioral experiments with the participants. One is that they reflect past biases and prejudices of the participants. So, in this respect, they do not model the Kuhnian revolutions of social norms (Indurkhya 2016). Consider, for instance, the work of Ni *et al.* (2011), who trained their program with the official UK top-40 singles chart over the past 50 years to learn as to what makes a song popular. A program like this might successfully predict the winner of the future Eurovision competitions, but it cannot predict drastic changes in the aesthetic values and tastes like atonal music or abstract art.

Another limitation is that once the algorithmic methods are known, people alter their behavior in order to achieve the desired result. I will refer to this as the *Minority Report* effect, for it was the basis of a short story with this title by Philip K. Dick. A case in point is the manipulation of electoral district boundaries in the US by individual parties in order to give them a demographic advantage, which is known as *Gerrymandering* (Mann 2006).

Conclusions: Assuming that the autonomous decision-making systems are here to stay, and that there will be situations in which they will be making moral and ethical decisions, in order that these decisions are accepted by many (if not all) humans, it is crucial to generate explanations underlying those decisions that are psychologically convincing. To emphasize, a rational or logical explanation is not always psychologically compelling. So even though a machine may make a decision based on some calculated probabilities based on logic, it is important to explain it from a psychological point of view. This is illustrated by the experience of the designers of one of the first expert systems *Mycin* (Shortliffe 1976), which was found to be lacking in explanations, and this feature was added later in *Emycin* (Ulug 1986).

More recently, the same concern was echoed by the head of Google's self-driving car project Dmitri Dolgov: "Over the last year, we've learned that being a good driver is more than just knowing how to safely navigate around people, [it's also about] knowing how to interact with them." (Quoted in Wall 2016). BBC Technology Editor, Matthew Walls notes: "Driving isn't just about technology and engineering, it's about human interactions and psychology." The same can be said about moral decision-making: it is not just about rationality and logic. To make moral decisions that can be supported by psychologically acceptable explanations, it is important to research how humans reason and what arguments they find persuasive, and incorporate this ability in robots and other autonomous systems. We have outlined an approach to model this aspect in our earlier research (Indurkhya and Misztal-Radecka 2016), and are working towards implementing these ideas.

References

Arkin, R.C., Ulam, P. & Wagner, A.R. 2012. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust and deception. *Proceedings of the IEEE 100* (3).

Bauman, C. W., McGraw, A. P., Bartels, D. M. & Warren, C. 2014. Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass 8,* 536 – 554.

Ben-Ary, Guy & Ben-Ary Gemma 2016. Bio-engineered brains and robotic bodies: From embodiment to self-portraiture. In D. Herath, C.Kroos & Stelarc (eds.) *Robots and art,* Singapore: Springer, 307 – 326.

Bruers, S. & Braeckman, J. 2014. A review and systematization of the trolley problem. *Philosophia 42*(2): 251 –169.

Brogan, J. 2016. Should a Self-Driving Car Kill Two Jaywalkers or One Law-Abiding Citizen? http://www.slate.com/blogs/future_tense/2016/08/11/moral_machine_from_mit_poses_self_driving_car_thought_experiments.html Accessed on 5 Sept. 2016.

Bryson, J. J. 2016. Patiency is not a virtue: AI and the design of ethical systems. *Proceedings of the AAAI Spring Symposium on Ethical and Moral Considerations in Nonhuman Agents,* Palo Alto, CA: AAAI Press, 202–207.

Dawes, R.M., Faust, D. & Meehl P.E. 1989. Clinical versus actuarial judgment. *Science 243*(4899), 1668 – 1674.

Indurkhya, B. 2016. A cognitive perspective on norms. In J. Stelmach, B. Brożek and Ł. Kwiatek (eds.) The Normative Mind, Kraków (Poland): Copernicus Center Press, 35–63.

Indurkhya, B. & Misztal-Radecka, J. 2016. Incorporating human dimension in autonomous decision-making on moral and ethical issues, *Proceedings of the AAAI Spring Symposium on Ethical and Moral Considerations in Nonhuman Agents,* Palo Alto, CA: AAAI Press, 226–230.

Iserson, K.V. & Moskop, J.C. 2007. Triage in medicine, part I: Concept, history and types. *Annals of Emergency Medicine 49*(3), 275 – 281.

Krauss, D.A. & Sales, B.D. (2001). The effects of clinical and scientific expert testimony on juror decision making in capital sentencing. *Psychology, Public Policy and Law 7*(2), 267-310.

Levy, D. 2007. *Love and sex with robots: The evolution of human-robot relationships.* New York: HarperCollins.

Lin, P. 2011. Introduction to robot ethics. In P. Lin, K. Abney & G.A. Bekey (eds.) *Robot Ethics: The Ethical and Social Implications of Robotics.* Cambridge (Mass.): MIT Press, 3–16.

Lin, P., Abney, K. & Bekey G.A. (eds.) 2011. *Robot Ethics: The Ethical and Social Implications of Robotics.* Cambridge (Mass.): MIT Press.

Linhares, A. 2014. The emergence of choice: Decision-making and strategic thinking through analogies. *Information Sciences 259,* 36 –56.

Litwack, T. R. 2001. Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law 7*(2), 409 – 443.

Mann, T.E. 2006. Polarizing the House of Representatives: How much does Gerrymandering matter? In P.S.Nivola & D.W. Brady (eds.) *Red and Blue Nation?: Characteristics and Causes of America's Polarized Politics.* Baltimore: Brookings Institution Press, 263 – 300.

Morton. L. 1960. The decision to use the atomic bomb. In L. Morton *Command Decisions*. Office of the chief of the military history of the army. Washington D.C. 493 – 518. http://www.dod.mil/pubs/foi/Reading_Room/NCB/361.pdf Accessed on 5 Nov. 2016.

Moskop, J.C. & Iserson, K.V. 2007. Triage in medicine, part II: Underlying values and principles. *Annals of Emergency Medicine 49*(3), 282 – 287.

Navarrete, C. D., McDonald, M. M., Mott, Michael L. & Asher, B. 2012. Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem". *Emotion 12*(2), 364 – 370.

Ni, Y., Santos-Rodriguez, R., Mcvicar, M. & De Bie, T. 2011. Hit Song Science Once Again a Science? Fourth International Workshop on Machine Learning and Music: Learning from Musical Structure, Sierra Nevada, Spain.

Nucci, E.D. 2013. Self-sacrifice and the trolley problem. *Philosophical Psychology 26*(5), 662–672.

Robertson-Steel, I. 2006. Evolution of Triage system. *Emergency Medical Journal 23*(2), 154 – 155.

Schwartzman, M. 2011. *See yourself sensing: Redefining human perception.* London: Black Dog Publishing.

Searle, J. 2011. Watson doesn't know it won on "Jeopardy". *Wall Street Journal,* 23 February 2011.

Shortliffe, E.H. 1976. *Computer-based medical consultations: MYCIN.* New York: Elsevier/North Holland.

Starr, S.B. 2014. Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review 66.*

Ulug, F. 1986. *Emycin-Prolog expert system shell.* Master's Thesis. Naval Postgraduate School, Monterey, California.

Wall, M. 2016. Would you bully a driverless car or show it respect? http://www.bbc.com/news/business-37706666, accessed on 21 October 2016.

Warwick, K. 2003. Cyborg morals, cyborg values, cyborg ethics. *Ethics and Information Technology 5*(3), 131 – 137.

Warwick, K. 2010. Implications and consequences of robots with biological brains. *Ethics and Information Technology 12*(3), 223 – 234.

Warwick, K. 2014. The cyborg revolution. *Nanoethics 8*(3), 263 – 273.

Tomorrow's Eves and Adams: What Yesterday's Science Fiction Could Tell Us about our Future with Companion Robots

Elizabeth Kinne
Assistant of Professor of Comparative Literature and English, The American University of Paris, Paris, France
ekinne@aup.edu

Abstract

This article briefly explores how fictional representations of companion robots and an ethics of vulnerability can help make ethical decisions regarding the use of sex robots.

Companion Robots and Human Vulnerability

Providing or expressing care and affection is one of the primary functions in our current conception of companion robots. These robots can be used in therapeutic contexts, engaging and monitoring the emotional and mental needs or states of the elderly or mentally ill, in familial domestic contexts providing entertainment by participating in affective interactions, or in the intimate sphere simulating the emotional and physical companionship of a romantic partner. While the apparent similarities of these contexts might not be visible at first, what has all too often been overlooked in the widely varying needs of these human participants is their commonly shared vulnerability. This fundamental characteristic of humanity, vulnerability, has been associated with susceptibility to harm and violence and is both reductive negative as Erinn Gilson argues in her recent book *The Ethics of Vulnerability: A Feminist Analysis of Social Life and Practice* (2013). Gilson engages the reader with a broader conception of vulnerability defined as a condition of potential in which one is open to being affected by the environment and by others, "an unavoidable feature of life and, as such, is not simply an opening to harm but an opening to all experience, negative, positive, and ambiguous" (24). It is the very condition which allows for relationship to exist. In the case of human-machine interactions, it would not seem to be a shared quality and it is debatable as to whether or not a shared vulnerability is technically possible or ethically desirable.

Are some users more vulnerable than others?

In recent considerations regarding companion robots, the vulnerability of the human interacting with the machine has often been defined by their capacity to know whether or not they are interacting with a machine or not and whether they accord sentience or not to the machine, criteria that most often apply to the elderly, the mentally impaired or to children, but any user can be susceptible to or have the desire to maintain this perception. Within Gilson's framework, any human user is a vulnerable user and ethical and moral consequences result. By these same standards, whether or not vulnerability could and should be extended to the machine itself depends on how we determine what it means to be "affected", raising the question of its potential status as moral patient. David Levy, in his 2007 book *Love + Sex with Robots*, describes machines in which this line could become blurred, extending Turing's test of machine intelligence to the emotional and consensual (if it looks like it is expressing emotion, it has emotions; if it says that it consents and behaves in a way that demonstrates consent, it is consenting). Whether these assertions are valid or a vulnerability Turing test would be of any moral and ethical utility remains to be seen; what is of interest here is expanding the notion of vulnerability to all human users of robotic companions so that we can ask, if not answer, better questions regarding the ethical implications of their use. What are at stake are the types of human subjectivity that could be created in scenarios in which vulnerability is a factor *sine quo non* that founding a relationality whose authenticity is questionable in these contexts. In an essence, what type of human subjects do we want companionate technology to give us the possibility of becoming?

Imagining feeling machines and…

For the moment, the realm of the feeling machine remains that of science fiction as only authors and screenwriters explore the possibility for artificial sentience and feelings and express the desirability of moral patiency being granted to these machines as a result. It could be hoped that this remains a pure fiction; the utility of these fictional considerations for this paper are their numerous depictions of human vulnerability and subsequent rejections of human vulnerability that interaction with a machine and not another human allows. This paper will be draw on examples from *He, She and It* (1991)by Marge Piercy and *Tomorrow's Eve* (1886) by Auguste Villiers de l'Isle-Adam, however science fiction films such as *Her* (2013), *Ex Machina* (2015), *Bladerunner* (1982)and *A.I* .(2001) continue to raise these same questions.

In her 1991 novel feminist speculative fiction author Marge Piercy retells the legend of the golem in a 2059 dystopian America in which environmental catastrophe has struck and large corporations have dictatorial control over the daily lives of their workers. Piercy introduces her readership to Yod, a robot designed to protect the small enclave of resisters that create him. Shira, a young woman responsible for socializing the robot, falls in love with him and much of the plot of the novel revolves around their unlikely relationship.

In *Tomorrow's Eve*, a young Lord Ewald seeks the assistance of Thomas Edison in order to replace the young but purportedly foolish woman with whom he has fallen in love, with a mechanized semblance of her more in keeping with his demanding standards in a companion. Edison is more than happy to oblige, sympathetic to his plight, yet tragedy strikes the misogynist Ewald at the end of the novel.

…the humans that love them

These two depictions of fictional characters seeking mechanical companionship, Shira and Lord Ewald, share experienced disappointments with their former human companions. Shira, having been deprived of her maternal rights by her ex-husband, seeks a companion who will help her to restore those rights and serve as a protector for herself, her son, and her community. Yod gladly obliges her by playing a paternal and protective role, ultimately sacrificing himself for the welfare of all. Lord Ewald is infatuated by Miss Alicia Clary for her beauty, yet finds what he calls her "soul" deficient. The arrival of the real Miss Alicia Clary at the end of the novel allows the reader to suppose that what might be at stake is not so much her intellect as her pragmatic refusal to submit to his whims. The highly conventional nature of the gender roles depicted in these works strike the reader, as do the ways in which the human protagonists desire robotic companions which allow them to exert control over their social circumstances, heightening their personal agency in order to diminish their self-perceived vulnerabilities. Both also explore the ways in which Yod and Hadaly (the mechanical Alicia Clary) may or may not be affected by their human companions. The feminist implications of the first work and the misogyny of the latter function as two ends of the spectrum of possibilities for companion machines; the ethical implications of robotic companions are neither clear cut nor simple.

The refusal of human vulnerability for the primary users of romantic companion robots could be likened to Gilson's considerations regarding the interstices of the ethics of vulnerability and pornography. Gilson posits that by refusing openness to others and embracing invulnerability a form of entrepreneurial subjectivity is created with ethically damaging consequences in which "responsibility for risk and for common human vulnerabilities is increasingly privatized rather than shared" (Gilson 98). We see how the scenarios for use imagined by David Levy envelop the human user in a closed sexual system in which individual desires are mirrored but not truly shared by a robotic companion. Concerns raised by Gilson are also echoed in Kathleen Richardson's work which establishes a connection between sex robots and the exploitative possibilities of sex work. What many imagine, and what science fiction tells us, is that companion robots are desirable objects not just for sexual gratification, but also as the means for establishing invulnerable forms of relationality from which risk is absent. However, this desire for invulnerability is what precludes authenticity in human relationships. These considerations could be extended to develop an expansive notion of vulnerability when considering human users of companion robots in general and the consequences of non-mutuality or artificial reciprocity.

References

Gilson, Erinn. 2013. The Ethics of Vulnerability: A Feminist Analysis of Social Life and Practice. New York: Routledge.

Levy, David. 2007. Love + Sex with Robots. New York: Harper Perennial.

L'Isle-Adam, Villiers de. 2001. Tomorrow's Eve. Trans. Robert Martin Adams. Urbana, Illinois: University of Illinois P.

Piercy, Marge. He, She and It. New York: Fawcett, 1991.

Richardson, Kathleen.2015. An Anthropology of Robots: Annihilation, Anxiety, and Machines. New York: Routledge.

Robotics as a new object of EU's 'ethical' attention: preliminary considerations concerning EP's draft Report on Robotics

Mihalis Kritikos
Law, Science, Technology & Society (LSTS)-Vrije Universiteit Brussels (VUB)
mihail.kritikos@gmail.com

Abstract

As technology is expected to become even more human centred, the potential for empowerment through the use of robotics is nuanced by a set of tensions or risks upon human safety, privacy, integrity, dignity, autonomy and data ownership. While much of the promise held in these technological innovations remains to be fully realised, the expansion of robotics into new areas of human interaction and activity is expected to be followed by a profound set of shifts in the way individuals perceive some fundamental concepts such as companionship or intimacy. The 'human-centered' turn in robotics technologies raises ethical questions already at the design phase as it involves the gathering and volunteering of data, and the involvement of lay people in experimentation with robotics for the programming of the necessary algorithms.

Research funders and policymakers in the field of science and technology at the EU level increasingly make use of socioethical factors as their basis for decision-making. At the EU level, ethics embeddedness is provided through the form of institutional structures that are acting as centers of ethical expertise. This is seen as part of the so-called responsible innovation narrative. Based on the author's experience with the EU's Ethics structures, the paper will examine the reasons behind the development of an ethics governance framework at the EU level and will provide a mapping of the main challenges associated with the gradual strengthening of the ethical component of EU's research policies.

It will then shed light on the operation of ad hoc research ethics committees created for the purposes of EU-wide ethical evaluations and will assess whether the process for the establishment of an EU-wide institutional framework for the responsible conduct of research indicates a tendency for the establishment of centralized Community ethical standards. By focusing on the operation of the EU Ethics Review Panels, the objective of the paper is to analyse the procedural approach towards research ethics followed at the EU level and the opportunities as well as the challenges that this entails especially for robotics.

The paper will then highlight the main points of the recently drafted report of the European Parliament in relation to the ethical and legal aspects of robotics and will present the arduous process for its formulation.

In view of the upcoming human-centered challenges, a governing/guiding framework for the design, production and use of robots is needed to guide and/or compliment the respective legal recommendations or even the existing national or EU acquis. The proposed framework takes the form of a code of conduct for researchers/designers and users, a code for ethics committees when reviewing robotics protocols and of 2 model licences for engineers and users. The framework will be based on the principles enshrined in the EU Charter of Fundamental Rights (such as human dignity and human rights, equality, justice and equity, benefit and harm, non-discrimination and non-stigmatization, autonomy and individual responsibility, informed consent, privacy and social responsibility) and on existing ethical practices and codes.

The values enshrined in the EU Charter of Fundamental Rights represent the normative framework on which a common understanding of the ethical risks associated with the operation of robots could be built. Still, judgements about the ethical soundness of robotics applications depend significantly on the specific context of application and the findings of the respective risk assessment process. As a result, the report has been strongly inspired by the stream of engineering ethics that places special responsibility on the engineers involved in the making of the machine with the focus on the moral decisions and responsibilities of designers.

Within this frame, the main tenets of the report - including the proposed (ethics-related) Magna Charta of Robotics) will be discussed and particular attention will be paid to the various ways the draft report is inspired and guided by the field of machine ethics and machine law.

Moreover, a detailed analysis of the legal backcasting process that was initiated in the frame of this drafting process will be provided. This reflection process was used for the first time as part of a foresight exercise conducted by the European Parliament. The overarching purpose of this 'legal backcasting' phase was to support the European Parliament (parliamentary committees and Intergroups), as well as the individual Members, to act proactively, when performing legislative work, in view of the rapid developments in the field of robotics.

This step of the foresight process, which resulted in briefings for the European Parliament aimed at translating the findings of the foresight phase in legal terms so as to pave the way for possible parliament reflection and work. This phase transformed the outcomes from the previous steps into a forward looking instrument for the European Parliament, the parliamentary committees and the Members of the European Parliament.

It consisted of the following phases:
1.-Identification and analysis of areas of possible future concern regarding CPS that may trigger EU legal interest;
2. Identification of those relevant EP committees and Intergroups of the EP that may have a stake or interest in these areas;
3. Identification of those legal instruments that may need to be reviewed, modified or further specified;
4. Identification of possible horizontal issues of legal nature (not committee-specific, wider questions to think about);

The analysis looked at the different ways in which the current EU legislative framework may be affected by advances in robotics and by the respective technological trends. To do so, a scanning of the current state-of-the-art of legislation pertaining to robotics was performed pointing towards mostly areas of EU law that are in need of adjustment or revision due to the initiation of emerging robotics technologies. The focus has primarily been on whether robots raise particular legal concerns or challenges and whether these can be addressed within the existing EU legal framework rather than on how human behaviour might be regulated through robotics.

The focus on the existing EU legal framework does not necessarily imply that all robotic applications by and large can be accommodated within the current boundaries of EU Law or that the adoption of a uniform body of law or of a single legal approach towards CPS as a whole (a form of lex robotica) should be excluded given the transnational character of some of these challenges.

Although the regulatory implications of robotics can be approached from a variety of legal perspectives, the legal analysis does not attempt to prejudge what will eventually be the most appropriate instrument in each case. For some types of applications and some regulatory domains, a review is recommended, while for some others, robotics can possibly be regulated by modifying existing directives or regulations following a case-by-case approach, international conventions or soft law approaches such as guidelines, codes of conduct, or standards drawn up by professional associations or technical standardisation organisations such as the International Organization for Standardization ISO or European organisations such as CEN and CENELEC.

Given the cross-sectoral nature of robotics as an object of ethical and legal inquiry, the paper pays particular attention to the constraints and safeguards that the draft report is planned to introduce so as to allow decision-makers and stakeholders to handle and eventually control tensions or risks upon human safety, privacy, integrity, dignity, autonomy and data ownership.

References

Anderson, M. and Anderson, S. L., eds. (2011). Machine Ethics.

Bosk, C.L. and De Vries, R.G. (2004) ‘Bureaucracies of mass deception: Institutional Review Boards and the ethics of ethnographic research’, Annals of the American Academy of Political and Social Science, 595, September: 249–263

EPSRC. (2011). Principles of robotics: Regulating Robots in the Real World (Tech. Rep.). Engineering and Physical Science Research Council. EPSRC

Guillemin, M. and Gillam, L. (2004) ‘Ethics, reflexivity and “ethically important moment”’, in research’, Qualitative Inquiry, 10: 261–280

Ingram, B., Jones, D., Lewis, A., Richards, M., Rich, C., and Schachterle, L. (2010). A code of ethics for robotics engineers. In Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI).

Lin, P., Abney K., Bekey G., (eds),(2014) Robot Ethics : The Ethical and Social Implications of Robotics, MIT Press.

Wallach, W. Allen, C. (2009), Moral Machines: Teaching robots right from wrong, Oxford/New York, Oxford University Press

Tallacchini, M., 2009, ‘Governing by Values. EU Ethics: Soft Tool, Hard Effects’, Minerva 47: 281-306.

Is Machine Ethics computable, non-computable or nonsensical?

Roman Krzanowski[1], Kamil Mamak[2], Kamil Trombik[3] and Ewelina Gradzka[4]
[1]Pontifical University of John Paul II, Cracow
[2]Pontifical University of John Paul II, Cracow
[3]Pontifical University of John Paul II, Cracow
[4]Jesuit University Ignatianum, Cracow
rmkrzan@gmail.com, kamil.mamak@gmail.com, kamil.trombik@gmail.com, ewelina.gradzka@yahoo.com

This essay argues that the question of computability of ethics in autonomous machines is a nonsensical one. It is nonsensical because the question attempts to endow a computer with some metaphysical qualities which, because of its nature, a computer does not have and considering what computing and ethics are, in high probability never will have (f. ex. Wallach, Allen 2009). The main theses of this paper are that:

- Confounding ethics with the performance of computing machines leads to a categorical error and misunderstanding of what computing machines are, what they do and what ethics is.
- Development of autonomous machines should concentrate on producing software that creates 'safe' machines not ethical automata, as this would imply qualities that machine cannot have.

Ethics, for some of us, is a set of prescriptions or rules on how to live a good and rewarding life as an individual and as a member of a society (Burke 2008; Vardy Grosh 1999; MacIntyre 1998). Such a concept of ethics may be reduced, as is often the case, to a set of rules specifying what to do. Such a set of rules, based on Hobbesian, Kantian, Mills or other ethical schools, can be to some extent translated into a computer algorithm; no assumptions of any 'metaphysical' dimension of the actor are assumed in this case. Then, if a machine is programmed according to these rules, one may claim that it possesses ethical qualities or is an ethical machine (Anderson Anderson 2010). But ethics is more than just rules. Ethics comes as a package deal. Ethics (implicitly or explicitly) requires free will, consciousness, a concept of good and wrong, an understanding of responsibility (of "oughts" and "oughts not") (MacIntyre 1998; Veach 1973, Sandel 2010), and some comprehension of reality around us. A lot of deep metaphysics is involved in the concept of ethics, such as free will, good life, or the individual. Dispensing with metaphysics leaves ethical statements groundless.

Autonomous machines (f. ex. Human Rights Watch 2015; Floreano at al. 1998; Patrick et al. 2008; Ni 2016), or machines that act in the environment without direct command or involvement from man, are all in essence computers. The difference between them and a laptop is that these machines can walk, fly, float, maybe talk, maybe move around, shoot, kill and do other things; in short they can interact with us. An autonomous machine may be a vacuum cleaner or it could be a computer in the guise of a beautiful woman, a drone loaded with deadly weapons, or a 10 ton truck.

All of these 'things' are programmed devices based in their core, as all computing devices are, on the Universal Turing Machine - or the UTM (f.ex. Bulker- Plummer 2011). The UTM is a theoretical device that processes strings of 0s and 1s, according to few simple rules, and not much more (f. ex. Feynman 2000; Evans 2011; Tucker 2004). Whatever rules of behavior we program into a computer or a computing device, they will always boil down to the enumerations of the sequence of the transitions rules of the state machine (Gowers 2002) in a form of 0s and 1s.

By ethics in autonomous machines we would understand a program or a set of rules driving the machine's behavior. But as any program in a computing device is in its essence "the sequence of the transitions rules of the state machine", ethics implemented in a computer device, or ' computer ethics', is no more or no less than just that.

What is non-computability ? In the most simplistic terms non-computability means that for a given task, or an input, the computing device cannot (in reasonable time or at all) reach the 'END' of a program; it keeps computing or gets stuck; a computer program, by definition, is a procedure that at some point will end, providing some results (Evans 2011). The non-computability of a program indicates that computer was given a task that cannot be computed; it is not the program that is wrong, there is nothing wrong with a computer itself. This is the very nature of some problems that they cannot be computed. Thus, non- computability in a way means that the non-computable program represents the problem that is not suited for a computer (f. ex. Cathcart 2013). What does non- computability mean for ethics? It means that ethic is one of these problems that are not suited for computing machines (f .ex. Moor 2006,).

One of the theses of this paper is that the problem of computability of ethics in computing machines should not have even arisen because it is, in principle, non-existent, or as we said nonsensical; it is so as we claim computing or computability and principles of ethics belong to two different conceptual realms having nothing to do with each other. What is possibly computable or not, is not ethics, it is a sequence of simple operations acted upon in a mechanical way in response to some, even complex, input. The output of such an ethical program can be regarded as an ethical decision only by crude equivocation and, only with the help of the same fallacy a machine with such an algorithm can be called an ethical machine (rather than a computing device with some behavioral rules).

The term machine ethics confounds what ethics is and what computing is. It would be more intellectually honest and semantically clean to talk about software and decision-rules embedded in machines rather than a machine with embedded ethics, what the term machine ethics implies. Confusing the meaning of terms will usually lead to gross misinterpretation of reality, with serious consequences, as for example Varoufakis shows in his analysis of the roots of the 2008 economic crash (Varoufakis 2015).

Without deep metaphysics we may produce an "ethical machine" akin to a psychopath (Zizek 2006), not an ethical individual. By a psychopath I mean an individual that makes logical decisions (using his logic) but not ethical. This points is however lost and research multiplies in which the term "machine ethics" or similar is used to denote the procedures controlling computing devices (f. ex. Wallach et. al. 2010; McDermott, 2008; Torrance 2008). Our task, the task of philosophers, engineers, scientists, should be to point out the misuse of terminology, so the confusion between what is what, and consequences of thinking with washed out terms, are minimized.

Disregarding differences between ethics and computing will inevitably lead in the most benign case, to categorical mix-up (upsetting no-one but philosophers) and in the worst case to the total confusion over what we are and what machines are. Blurring of the boundaries between us and artifacts creates a potentially poisonous admixture of ideas: calling a set of programming rules running on an autonomous computer "ethics" bestows on it, by virtue of association, all that comes with ethics: moral responsibility, moral stature, maybe even free will. Eventually, we assign to such a machine a personhood and all the rights and responsibilities coming with it (absconding in the same time from any responsibilities for what these machines do); we would call it 'machine antropomorphisation'.

For autonomous machines the real question is not that of computability of some ethical rules, but whether we can develop a program that could prevent machines from harming us, so we will not become the victims of our own creation (f. ex. Moor 1979) ; Sci-fi literature abounds with such dark scenarios (f. ex. Bostrom 2015; Lem 2013).

Several objections to the argument in this paper are possible. One may claim that the term 'computability' may mean not more than just 'solving a problem' or 'making a decision'. Thus computability of ethics would mean just making ethical decision. Such a use of the term, however, would actually change the meaning of 'computability' from the precise concept related to the essence of computing to a poetic metaphor belonging rather to the realm of literature than technical discourse, thus making any discussion of computability of ethics meaningless.

One could also claim that 'computing' is in fact an act of thinking. But this would make a concept of computing even more nebulous, as we still are not sure what is the essence of thinking,; not to mention that noncomputability would be rather an odd concept here (imagine noncomputability of thinking).

One may also point out that using the UTM paradigm for computing is too limited or restrictive and bound to the 'current state of art'. This is obviously true in a sense that any computing device we know of now can be reduced to the UTM concept. However, we do not know any other paradigm of computing and none is on the horizon. So, as far as we can see, as no new paradigm of computing is emerging, the current state of art is for the time being what the future one will be as well.

One can object to the comparison of computers to the UTM which is a theoretical concept, not an engineering one, thus having, one may argue, no import into practical issues. But by the same token one can object to the modeling of the gasoline engine by the Carnot principle as the principle is the theoretical one and nowhere visible in the engine.

One could propose that the problems raised in the paper are the results of some linguistic misunderstanding, we just use wrong words when we talk about machine ethics. I sense here a linguistic longing that all we have to do to solve all our philosophical problems is to use the proper and clear terminology (use the proper language). And in a sense it is true. But it would not get us into the essence of the problem as the problem presented here is about confusing concepts rather than misuse of words. Thus, assuming that our problem is the 'linguistic turn' we would just 'pass the buck' but not get into the heart of the matter.

I would rather refuse to consider in this paper any ' if-' scenarios borrowed for example from Blade Runner or Star Trek. They seem to be more at home in sci-fi literature than in a philosophical paper, and therefore should not have any bearings on the presented analysis. I would end the paper with the quotation by Susan

Schneider “When it comes to AI, philosophy is a matter of life and death.” (Conn, 2016). Let us bear this in mind when talking about ethical machines.

References

Anderson, M., S.L. Anderson. 2010. *Robot be Good*. Scientific American, 10,: 72-77.

Anderson, J. M., Kalra, N., Stanley, K. D. 2016. Sorensen, P., Samaras, C., O. Oluwatob. *Autonomous Vehicle Technology. A Guide for Policymakers.* Santa Monica: RAND Corporation.

Arkin, R. 2016. *Governing Lethal Behavior, Embedding ethics in a Hybrid Deliberative/Reactive Robot Architecture.* Technical Report GIT-GVU-07-11. Accessed on 02.06.2016, available on line at http://www.cc.gatech.edu/ai/robot-lab/online- publications/formalizationv35.pdf.

Beavers, A. F. 2011. *Is Ethics Computable.* Presidentiall Address- Aarhus, Debmark, July 4th. On line http://faculty.evansville.edu/tb2/PDFs/IACAP%202011% 20Presidential%20Address.pdf. accessed on 07.07.2016.

Blade Runner. 2016. Entry in Wikipedia. Accessed on 08.082016 at https://en.wikipedia.org /wiki/ Blade_Runner.

Bostrom, N. 2015. *Superintelligence.*, Oxford: Oxford University Press. Bourke, V. J. 2008. History of Ethics, Vol.I, V.II. Mount Jackson: Axios Press.

Barker-Plummer, David, *"Turing Machines"*, The Stanford Encyclopedia of Philosophy (Spring 2016 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford .edu/ archives/ spr2016/entries/turing-machine/

Cathcart, T. 2013. *The Trolley Proble.* New York: Workman Publishing, New York.

Conn, A. 2016. *The Ethical Questions Behind Artificial Intelligence*. accessed on 29.10.2016, available at http://www.huffingtonpost.com/entry/the-ethical-questions-behind-artificial-intelligence_us_ 580e9a2fe4 b0b1bd89fdb6e 9?timestamp =147735 3327956&utm _content= buffer19b3 8&utm_medium =social &utm_ source=facebook.com&utm_campaign=buffer

Edgar, S. T. 2003. *Morality and Ethics.* Boston: Jones and Bartlett Publishers. Evans, D. 2011. Introduction to Computing. Creative Commons.

Feynman, R. P. 2000. *Feynman Lectures On Computation.* New York: Westview Press.

Floreano, D., Godjecac, J., Martinoli, F., and J-D. Nicoud. 1998. *Design, Control and Applications of autonomous mobile robots.* Swiss Federal Institute of Technology, Lausanne. 1998, accessed on 2.08.2016, available at https://infoscience.epfl.ch/record/63893/files/aias.pdf. Gowers, T. 2002. *Mathematics. A very short introduction*. Oxford: Oxford University Press..

Lem, S. 2000. *Cyberiad*. London: Mariner Books. MacIntyre, A. 1998. *A Short History of Ethics*. Notre Dame: Notre Dame Press.

McDermott, D. 2008. *Why ethics is a high hurdle for AI.* Paper presented at 2008 North American Conference on Computing and Philosophy. Bloomington, Indiana.

Mind the Gap. 2015. Human Rights Watch.

Moor, J. H. 2006. *The Nature, Importance and Difficulty of Machine Ethics.* IEEE Intelligent Systems. July/August: 18-21.

Moor, J. H. 1979. *Are There Decisions Computers Should Never Make.* Nature and Systems. 1: 217-229.

Ni, R., Leug. J.2016. *Safety and Liability of Autonomous Vehicle Technologies.* Accessed on 06.06.2016. Available at https://groups.csail.mit.edu/ mac/classes/6.805 /student-papers/fall14-papers/Autonomous_ Vehicle Technologies.pdf .

Patrick,L., Bekey, G. Abney, K. 2008. *Autonomous Military Robotics: Risk, Ethics, Design.* US Department of Navy, Office of Naval Research.

Rappaport, W. 2016. *Philosophy of Computer Science*. Bufflo: University at Bufflo.

Reynolds, C. J. 2016. *On the computational Complexity of action evaluation.* On line http://affect.media.mit.edu /pdfs/05.reynolds-cepe.pdf, accessed on 07.07.2016.

Sandel, M.J. Justice. W*hat’s the right thing to do?* Penguin Books, London 2010.

Sellar S., J. Hosper. 1970. *Readings in Ethical Theory*. New York: Appleton-Century-Croft. Star Trek. accessed on 2.08.2016, available at https://pl. wikipedia.org/wiki/Star_Trek.

Torrance, S. 2008. *Ethics and consciousness in artificial agents.* AI and Society. 22: 295-521.

Tucker, A. B. 2004. *Computer Science Handbook.* Boca Raton: CRC Press.

Veach H. B.. 1973. *Rational Man.* London: Indiana University Press.

Vardy P., Grosch, P. 1999. *The Puzzle of Ethics.* London: Fount.

Varoufakis, Y. 2015. *Global Minotau.* London: Zed Books.

Wallach, W., Allen, C. 2011. *Moral Machines*. Oxford: Oxford University Press.

Wallach, W. Franklin, S., Allen C. A 2010. *Conceptual and Computational Model of Moral Decision Making in Human and Artificial Agents.* Topics in Cognitive Science, 2: 454–485.

Williams B. 1988. *Ethics and the Limits of Philosoph.* Havard: Harvard University Press.

Zizek, S. 2006. *How to read Lacan.* London: Granta Books.

Machine Ethics in the Context of the Internet of Things Environment

Ewa Łukasik

Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland
Ewa.Lukasik@cs.put.poznan.pl

Abstract

The Internet of Things (IoT), where objects of the physical world and the information world are capable of being identified and integrated into communication networks and become autonomous agents put new constraints on the responsibility of humans and machines. The paper discusses current awareness of this problem from the perspective of the IoT and Machine Ethics.

Introduction

It is demanded that an intelligent autonomous agent should make good decisions, based on available data even if the data is uncertain, missing, noisy or incorrect. Investigating the question of good and wrong machine actions and also machine responsibility, one has to take into account the environmental infrastructure in which the autonomous machine operates. Currently this is the Internet of Things (IoT) i.e. the global system where the ubiquitous elements of the physical world equipped in sensors are interconnected via the Internet (Ashton, 2009). The forecast is (Greenough, 2016) that by 2020 there will be 34 billion devices connected to the Internet (triple as many as in 2015), from which only one third will be smartphones tablets and watches. Being distributed in various locations *things* activate communication generating *big data* that, with the development of the system, will become extremely difficult to be efficiently and automatically managed, analyzed and understood. Equipped with intelligent software agents they will make intermediate decisions that other agents will rely on and use for their decisions.

Motivation

The standardization body, ITU-T, sets some requirements on the IoT system, including the interoperability among heterogeneous and distributed systems, autonomic networking (self management self-configuring, self-healing, self-protecting) and autonomic services provisioning (ITU-T Recommendation, 2012). The convergence of natural environment with technology will lead to the creation of hybrid ecologies, where the responsibility for both humans and autonomous machines (robots, autonomous vehicles etc.) will be highly interdependent. Operations will rely on a multitude of data, decisions and services of distant, often unknown sources and may cause possible unethical decisions.

Aim

The aim of the contribution is to identify the ethical issues of the Internet of Things environment in the scope of machine ethics realm and to recognize lines of research in this domain.

Methods

First the technical aspects of the IoT will be considered. Then the morally relevant views on the IoT, as discussed in the literature, will be presented and further opposed to sample research in Machine Ethics.

Ethical issues related to the Internet of Things

Current perception of the Internet of Things.

The Internet of Things is becoming grounded in the technology, business and the human consciousness. The research is directed towards approaches that allow things to become smarter, more reliable and more autonomous, simply – more intelligent (Kyriazis, 2013). New architectures are proposed permitting things to learn from experience of the others (Kyriazis, 2013) capturing social behavior and preserving privacy (e.g. by privelets). Smart devices provide a basis for smart homes, smart cities, smart cars etc. The constantly evolving IoT requires continual software adaptation and this is moving towards mobile agents application (Fortino, 2016, Mzahm, 2014). Autonomic *things* will allow systems to self-manage the complexity, to control the dynamicity of growth and the distribution of the IoT (Leppänen, 2014). They will evolve to create the knowledge out of data and rules discovered during the operation.

The main concern of the IoT ecosystem architects is the cybersecurity best practices, e.g. network, endpoints and mobile device protection, data in motion and data at rest defenses as well as analysis and correlation tools (Vormetric, 2015). Also the need for privacy protection support and equal access are addressed, especially in high quality and highly secure human body related services (ITU-T Recommendation, 2012).

As the Internet of Things infrastructure and services grow, human opinion is driven towards the benefits that the technology brings, e.g. utility, well-being, sustainability, health, safety and security. However it is still unclear how Internet of Things is going to affect global trends across all spheres of human existence.

Morally relevant aspects of the Internet of Things

Moral arguments in favor of the Internet of Things are also accompanied by those raising its dangers and their possible preventive measures. Authors (e.g. Popescu and Georgescu, 2013, Ebersold and Glass, 2016) often recall Wachtel Report of EC meeting (2012) and Van Den Hoven (2014), where 11 defining features of IoT were characterized causing the ethical problems, e.g.:

- ambiguous criteria of identity and system boundaries because of an easy transformation of natural objects, artefacts and human beings,
- electronic identity of objects with various levels of importance, crucial for IoT, but difficult to be managed; even if not maliciously used, they may be simply wrongly managed or erroneous,
- unprecedented degree of connectivity between objects and humans in networks (Connectivity)
- spontaneous and unexpected (for users and designers) interference of interconnected objects driven by autonomous agents,
- objects with embedded intelligence will make humans feel cognitively and physically handicapped; some will not accept the embodiment of *extended mind*.

There are concerns about the distributed control and governance of IoT that will be faced with the unpredictable problems and uncertainty in which neither human, nor an autonomous machine will have relevant knowledge to make right/ethical decisions.

The Van Den Hoven group (2012) expects the remedial response for those constraints in Value Sensitive Design, i.e. Values Built into Systems and Responsible Design of Socio – Technical Systems. Engineers are Choice Architects that design for X, where X is e.g. privacy, inclusion, sustainability, democracy, safety, transparency, accountability, human capabilities. However it is still unknown, how to put these ideas into life.

Context of the research in Machine Ethics.

There is a multitude approaches for and against artificial morality and as many proposals for solving the problem of machine ethics or trying to answer whether artificial moral agents are computationally feasible. The IoT constrained features impose considerable difficulties on modeling behavior of ethical agents. Whether they are rule or data driven – their behavior, in author's opinion, can be hardly predicted in a complicated ecosystem of the IoT. Howard and Muntean (2016) propose a model for an artificial autonomous moral agent (AAMA), which is minimal in its ethical assumptions. Starting from a set of moral data, AAMA is able to learn and develop a form of moral competency. As a drawback, the authors see the dependency on the data, their reliability, or the way they were collected.

Some clues for approaching the problem of machine ethics in IoT environment can be taken from Floridi and Sanders seminal paper (2004) where the Method of Abstraction for analyzing the level of abstraction (LoA) at which an agent is considered to act is proposed. The LoA is determined by the way in which one chooses to describe, analyze and discuss a system and its context. The moral agenthood, depends on a LoA. This approach was criticized by Grodzinsky and Miller (2008).

The radical view was presented by Hew (2014), who claimed that "with foreseeable technologies, an artificial agent will carry zero responsibility for its behavior and humans will retain full responsibility" and Deng (2015) concluded "We need some serious progress to figure out what's relevant for artificial intelligence to reason successfully in ethical situations".

Computer scientists trying to respond to this question used to rely mainly on logic rules. This approach may work in static circumstances, but taking decision in a dynamically changing ecosystem is much more complicated and much less predictable (ibid.).

Conclusions

We have presented some issues related to the problem of Machine Ethics in the dynamically evolving environment of the Internet of Things, which, because of its complexity, dependability, unpredictability and dynamics will be hard to manage as a whole by both humans and machines. For the same reasons ensuring ethical action or delegating ethical responsibility to any entity is intricate and abstruse. Nevertheless the constant research efforts should be made in this domain together with promoting awareness of ethical risks from the Internet of Things among researchers engineers and students. As for now, the knowledge of those problems is very limited.

References

Ashton K. 2009. That 'Internet of Things' Thing. *RFiD Journal*, June. http://www.rfidjournal.com/articles/view?4986

Deng B. 2015. Machine ethics: The robot's dilemma. *Nature*, vol. 523 issue 7558.

Ebersold K., and Glass R. 2016, The Internet of Things: a cause for ethical concern, *Issues in Information Systems*. Vol. 17, Issue IV, 145-151.

Floridi L., and Sanders J.W. 2004. On the Morality of Artificial Agents, *Minds and Machine* 14, 349–379.

Fortino G., Russo W., and Savaglio C. 2016. Simulation of Agent-oriented Internet of Things Systems. In Proc. 17th Workshop "From Objects to Agents". 8-13.

Greenough J. 2016. How the Internet of Things will impact consumers, businesses, and governments in 2016 and beyond, *Business Insider*, July. www.businessinsider.com.

Grodzinsky F.S., Miller K.W., and Wolf M.J. 2008. The ethics of designing artificial agents, *Ethics and Information Technology*. 10 (2-3), 115-121.

Hew P. Ch. 2014. Artificial moral agents are infeasible with foreseeable technologies, *Ethics and Information Technology*. vol. 16, issue 3. 197–206.

Howard D., and Mountean I. 2016. A Minimalist Model of the Artificial Autonomous Moral Agent (AAMA). In *AAAI Spring Symposium Series*.

ITU-T Recommendation Y.2060. 2012. Overview of the Internet of Things, 06.

Kyriazis D., and Varvarigou T. 2013. Smart, Autonomous and Reliable Internet of Things, *Procedia Computer Science*, vol. 21, 442-448.

Leppänen T., Riekki J., Liu M., Harjula E. and Ojala T. 2014. Mobile Agents-Based Smart Objects for the Internet of Things. In Fortino G. and Trunfio P. (eds.): *Internet of Things Based on Smart Objects*: Springer, pp. 29-48,

Mzahm A. M., Ahmad M.S., and Tang A.Y.C. 2014. Enhancing the Internet of Things (IoT) via the Concept of Agent of Things (AoT), *Journal of Network and Innovative Computing*. Vol 2. 101-110.

Popescul D. and Georgescu M. 2013. Internet of Things – some ethical issues. *The USV Annals of Economics and Public Administration.* Vol. 13, Issue 2(18). 208-214.

Van Den Hoven J. 2012. Ethics and The Internet of Things. European Commission. Delft University of Technology, http://ec.europa.eu/transparency/regexpert/index.cfm?do

=groupDetail.groupDetailDoc&id=7607&no=4.

Van Den Hoven, J. 2014. Fact sheet- Ethics Subgroup IoT – Ver. 4.0, http://digitalchampion.bg/uploads/agenda/en/filepath_85.pdf.

Vormetric Global Insider Threat Report, 2015.

Wachtel T. 2012. *IoT Expert Group Final Meeting Report*. European Commission, 14 November 2012.

Criminal liability of autonomous software. Critique of Gabriel Hallevy’s conception

Rafał Michalczak

Jagiellonian University
michalczak.rafal@gmail.com

Abstract

The discussion (especially outside the academia) about autonomous vehicles heats up. Many articles discussing the issue of proper design of machines which are able to kill people have been recently published. The questions were asked e.g.: who should they rather kill (assuming they have choice) [Dvorsky 2016] or what kind of ethics should guide them [Knight 2015]. Of course those ethical questions have their legal counterparts e.g.: how to design a policy concerning autonomous software or how (and to whom) ascribe a responsibility for the actions of autonomous software.

Many of those questions pass by rather unnoticed especially in the domain of the civil (as opposing to criminal) law. The autonomous software is here but the norms of the civil law haven’t change so much. The software is treated as a tool and the responsibility or liability for its actions is distributed accordingly. The issues of stock trading and concluding contracts don’t seem to interest the public opinion. But, as stated before, the approach differs in the case of “killer robots”.

From the perspective of academic inquiry the situation is quite opposite [Pagallo 2013]. Many works on the liability [Čerkaa, Grigienėa, Sirbikytėb 2015], the contract conclusion [Balke, Eymann 2008], [Allen, Widdison 1996], or the authorship [McCutcheon 2012] exist. However the investigations of how to design the criminal responsibility in peaceful situations where autonomous software is engaged, are rather scarce. (The state of affairs is different in the domain of war situations [Arkin 2009]).

Nevertheless I would like to examine a profound project of a design of a criminal responsibility put forward by Gabriel Hallevy [2013, 2015]. Although I find the project very inspiring I will try to show, that some analogies assumed in Hallevy’s ideas are not fully convincing. I will focus on the internal critique. I will not discuss some external arguments aimed at the very idea of criminal responsibility of autonomous software [Cevenini 2004].

Hallevy’s claims may be divided into two categories. The first one is about the ascription of criminal responsibility to autonomous software. The second one is about punishing those autonomous software. In both categories Hallevy bases his investigations on analogy between criminal responsibility of legal persons and criminal responsibility of autonomous software. The proposition is, in its general framework, very interesting and may be considered quite practical. However I think there are two main discrepancies which may undermine assumed analogy between legal persons and autonomous software.

The first one concerns the issue of ascription of criminal responsibility. The problem, I will investigate, reflects the ontological difference between legal persons and autonomous software. Legal persons are held responsible for the actions initiated (and causally linked) by natural persons. The situation is different in the case of autonomous software, which actions are problematic not because of deeds of some natural person (e.g. user or creator). Those actions may be legally relevant *per se*, due to autonomous activity of the software. Whilst legal persons and autonomous software resembles each other as artificial human creations, the nature of its actions differs. In this manner the autonomous software seems more similar to natural beings than to artificial beings.

Second discrepancy I will examine concerns the area of punishing autonomous software. Hallevy claims, that punishments applicable to natural persons are translatable not only to punishments for legal persons but also to punishments for autonomous software. In my opinion it once more neglects the ontological difference between legal persons and autonomous software. Assuming that the main rationale for punishing artificial beings is functional, the above-mentioned simple translation isn’t fully justified. Modification of the future activity of software (which is the aim of punishment within the functional framework) seems attainable by completely different means, than in the case of legal (as well as natural) persons. The software has a distinctive feature of being easily reprogrammable. This important difference isn’t though implicitly encompassed in the Hallevy’s model.

The Hallevy’s model is a great starting point for investigation of the possible design of criminal responsibility of autonomous software. However, in my opinion the simple and elegant solutions it provides aren’t always fully warranted due to fact, that they dismiss the important difference between legal persons and autonomous software.

References

Allen, T., Widdison, R. 1996. Can Computers Make Contracts? *Harvard Journal of Law & Technology* 9(1). 25-52.

Arkin, R.C. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton, London, New York: CRC Press.

Balke, T., Eymann, T. 2008. The Conclusion of Contracts by Software Agents in the Eyes of the Law. *AAMAS '08 Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems* 2. 771-778.

Čerkaa, P., Grigienėa, J., Sirbikytėb, G. 2015. Liability for damages caused by artificial intelligence. *Computer Law & Security Review* 31. 376-389.

Cevenini, C. eds. 2004. *The Law and Electronic Agents: Proceedings of the LEA 04 workshop*. Bologna: GEDIT Edizioni. 178–9.

Dvorsky, G. 23.06.2016. *Your Self-Driving Car Will Be Programmed to Kill You—Deal With It*. URL: http://gizmodo.com/your-self-driving-car-will-be-programmed-to-kill-you-de-1782499265.

Hallevy, G. 2013. *When Robot Kill. Artificial Intelligence Under Criminal Law*. Boston: Northeastern University Press.

Hallevy, G. 2015. *Liability for Crimes Involving Artificial Intelligence Systems*. Cham, Heidelberg, New York, Dordrecht, London: Springer.

Knight, W. 29.07.2015. *How to Help Self-Driving Cars Make Ethical Decision*s. URL: https://www.technologyreview.com/s/539731/how-to-help-self-driving-cars-make-ethical-decisions/.

McCutcheon, J. 2012. The Vanishing Author In Computer-Generated Works: A Critical Analysis Of Recent Australian Case Law. *Melbourne University Law Review* 36. 915-969.

Pagallo, U. 2013. *The Laws of Robots. Crimes, Contracts, and Torts*. Dordrecht, Heidelberg, New York, London: Springer.

The legal outcome of recognizing and granting personhood rights to artificial intelligences and digimmortalized persons.

Kamil Muzyka

Instytut Nauk Prawnych Polskiej Akademii Nauk

Abstract:

The greatest problem in modern legislature is the lack of foresight, and being enacted mostly in hindsight. There are several exceptions from this rule, such as UNCLOS of 1982, or Charter of Fundamental Rights of the European Union, yet still the technology is outpacing legislature, like in the field of Drones or selfdriving cars. This also happens in the field of artificial intelligence and autonomous robotics law. Currently no country or international organization recognizes the issue of autonomous robots, artificially intelligent learning systems or effects of a whole brain emulation. As far as academic research in legal theory goes, such works on possible legal recognition and capacity for AIs has been published by various scholars since 1980's, with many approaches and mixed results.

The development of more complex neural networks, deep learning systems, cognitive simulation, logical evaluation programs, is gaining pace with more investments form private investors, and governmental or international scientific programs.

There are grave concerns about the legal capcity of an automous system and unit. Most prior concerns were based on the lethality of and possible criminal activity conducted by such system. Most current concerns revolve around the issues of privacy, transparency, accountability and dual use. The issue of the distribution of rights and responsibilities in a legal environment, where the diffusion of responsibility and liability occurs between human and non-human actors, especially when employing the industrial or machine internet or if the programming is the work of several unrelated entities.

Similarly, the introduction of legal autonomy and partial legal capacity might create possibilities in order to bypass the law, whilst the owner or the contractor to the unit might try to cover behind a substitution principle, granting the legal autonomy, capacity and liability to the unit, therefore dodging the any civil or criminal liability for the actions performed by the AI or robotic unit.

Additionally, one must recall that autonomy in decision making and activities such as recognition and route planning, doesn't involve legal reasoning. It is more difficult to teach an AI system or an autonomous robot the legal code, than to code certain behaviors into it's programming. Otherwise the rules to which the units must abide might be causing errors, resulting in damage to property or loss of lives.

However, the increasing number of AI entities in the fields such as auditing, judicial system, accounting or strategical decision making might not actually eliminate humans from the workforce, but keep them in a supervisory role, as being in charge of the transparency and accountability of the AI agents.

On the other hand not all approaches to AI and robot autonomy are based on man-made algorithms or machine learning. There are a lot of theoretical as well as scientific approaches in order to create a biomimetic "bionic" intelligence. While the whole brain emulation is still the domain of future studies and science fiction, there had been several experiments following the path to create a full real time simulated nervous system, in order to study it's behavior, interactions with the environment, and in the longer perspective to be the basis for creating novel neurocognitive treatments for severe medical conditions,

Furthermore, the emulation of a roundworms neural system into a running software for an experimental robotic body should be seen as a next step on the path to digitally preserving a human brain. While an uploaded brain of a canine might create more ethical than legal concerns, the legal aspects of emulating a human being into a form of a software might lead to either fulfillment of a transhumanist dream, or an incarnation of a cyberpunk nightmare. In the latter scenario, the lack of legal recognition for one's humanity, after the emulation process, may lead to the objectification of a former person. This may lead to instances of forcing one to transfer one's rights to another party, in exchange for the vague promise of being allowed to live free from any harmful interference, as copyright protected data, not an actual human being. Thus the two topics, the intelligent artificial agents and digimmortalized human beings, join in the case of exploitation of intelligent persons, that are void of any rights and protection, by current legislature. That is the legal split between the libertarian extropian transhumanism, and the technoprogressive movement, for the latter would be

less eager to treat AIs, autonomous robotic persons or hiveminds equals, rather using them as slaves to benefit humanity, by freeing it from the burden of labour.

The paper will look into several cases of pros and cons, of AI personhood and Digimmortalized rights, taking into consideration the problems of physical and virtual abuse, criminal law, commerce law, the loopholes in the US maritime law, privacy, surveillance. It will also look further in problems of applying the broadened catalogue of persons into the law as well as several compromise models.

Keywords: Artificial Intelligence, Law, Whole Brain Emulation, Personhood, Digimmortalization

References:

http://ieet.org/index.php/IEET/more/muzyka20130829 - as an IEET Intern, as well as other AI related articles
https://www.degruyter.com/view/j/jagi.2013.4.issue-3/jagi-2013-0010/jagi-2013-0010.xml
http://www.ibuk.pl/fiszka/136034/czlowiek-w-relacji-do-zwierzat-roslin-i-maszyn-w-kulturze-t-2-od-humanizmu-do-posthumanizmu29-zarys-prawa-w-odniesieniu-do-sztucznych-inteligencji-i-osob-emulowanych.html - The autor of the article included in this monography.
http://www.maska.psc.uj.edu.pl/varia/artykuly?p_p_id=56_INSTANCE_CVGinNlIwZo9&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-3&p_p_col_pos=1&p_p_col_count=2&groupId=40768330&articleId=46948958
PhD candidate at the institute for Legal Studiess of the Polish Academy of Sciences
https://freelab2014.wordpress.com/kamil-muzyka-mamo-dziadek-znowu-sie-zawiesil/

Law-abiding by Design: Legal Status and Regulation of Artificial Agents

Przemysław Pałka
PhD Researcher, Department of Law, European University Institute, Via Bolognese 156, 50139 Firenze, Italy
przemyslaw.palka@eui.eu

This article addresses the question: how to ensure that artificial agents act in accordance with the law? It is concerned with the incentive design and the form of the law, not its substance. It is argued that in order for artificial agents to act lawfully, they need to have a legal (normative) component in their source code. This requires both adequate legal incentives for the artificial agents' developers/operators to include such a component, as well as transformations in the form of law and legal databases, to make legal norms understandable by the machines.

On the factual level, two basic observations should be made. Firstly, artificial agents (AAs), i.e. both software agents and robots, Chopra and White (2011.), currently undertake actions traditionally performed by human beings (buying, selling, processing data, taking decisions etc.); and these actions are not legally neutral, i.e. it is conceivable that artificial agents infringe the law while undertaking them, Pagallo (2013.), Hallevy (2013.). Secondly, the law has traditionally been, and still is, designed assuming that the entities whose behavior it regulates are human beings, capable of knowing the law and making a decision of whether to follow it. Artificial agents, on the other hands, will have these capabilities only if they are designed with such an ability. Proposals on how to convince the developers to do so and to how facilitate this process make up the contribution of this paper.

Theoretical and Regulatory Challenges

The emergence of artificial agents, on the most general level, gave rise to two classes of legal challenges: theoretical and regulatory. The former, concerned with understanding, could be summed up with the question: is the existing conceptual legal framework adequate to talk about artificial agents, and if not, how should it be amended? The latter, functional in nature, could be summed up with the question: is any regulation on the side of law necessary, and if yes, how to regulate? As in many other parts of the law and new technologies scholarship, many legal authors tend to rush to the regulatory sphere. However, any regulatory claim is necessarily based on some, often implicit, (mis)understanding. For that reason, if the prescriptive regulatory claims are to be operable, theory should be right in the first place.

The theoretical challenge is therefore addressed first. On the meta level, the functions of legal concepts are explained (referring to the reality, convening information about law's factual assumptions and the content of norms, Sartor (2009a.)), the origin of their meaning addressed, the potential dangers of 'stretching' the concepts enumerated, and a method of creating new ones proposed, Palka (2016.). On the substantive level, available legal concepts, together with the scholarly contributions endorsing them, are critically surveyed.

The regulatory challenge is addressed based on examples from three fields of law: personal data protection, unfair commercial practices (advertising) and discrimination in access to goods and services.

The Extreme Views: Mere Tools vs. Persons

Two extreme views are present in the legal literature: the 'personalization' approach and the 'mere tools' approach. According to the former, artificial agents either meet or, more often, potentially could meet conditions to be treated as autonomous subjects of rights and obligations and could/should be granted the status of legal persons, Solum (1992.). According to the latter, artificial agents are tools like any other, and the role of law should be to clarify the liability rules for artificial agents' actions, Sartor (2009b.). The author of this article finds both approaches suboptimal, though for diverging reasons.

The 'personalization' approach is often based on confusion about both the characteristics of artificial agents, and the content of the legal concept of a person. Additionally, it does not propose any operable solutions to the problem of artificial agents' potentially infringing the law.

The 'mere tools' approach, on the other hand, is a necessary first step – liability rules should be clear – but proves insufficient. It is based either on the assumption that clarifying liability rules is enough to ensure that artificial agents' developers and/or users will take necessary steps to ensure that AAs actions conform with the law; or on the assumption that the role of law is to punish those who infringe it, instead of ensuring that it is not being infringed in the first place.

This article argues that the first assumption (factual) is not true, while the second assumption (axiological) is undesirable. Instead, it is argued, the regulation should also concentrate on the design of AAs, in order to ensure that they are designed in a way that prevents them from breaking the law.

The article consists of four sections.

Section One: Where Are We? In Facts and in Books

This section provides a brief overview of the characteristics of artificial agents, their categorization (primarily autonomous and automatic ones), the roles they already play in socio-economic life, and the legal problems that their actions give rise to. It also surveys the state of the art in the legal literature.

Section Two: Why Exactly Personification of Artificial Agents Does Not Solve Any Problem?

This section takes up the question: could/should artificial agents be granted the status of legal persons? Even though, according to the author, the negative answer is quite straightforward, there is a value in explaining why exactly this is so. In order to answer this question, the legal concept of a 'person' is reconstructed from the statutory and the doctrinal legal discourses and compared with the characteristics of AAs. The concept is reconstructed on the rules level (persons maximize utility, can err, are driven by emotions etc.) and on the meta-level (persons can know what the law is, are capable of applying the law to a given factual situation and taking a moral decision on whether to follow or infringe it).

The reconstruction of this concept, especially on the meta-level, proves helpful in explaining how the form of the law needs to be changed in order to ensure the AAs' action conformity with the law.

Section Three: Why Ex-Post Policing of AAs' Actions Proves Insufficient

This section addresses the question: is clarifying the rules on liability for AA's action sufficient to guarantee that their developers/operators take all the necessary steps to ensure that AAs act in accordance with the law? The negative answer is given, based both on the empirical data, Sweeney (2013.), and on the law and economics analysis of potential costs and benefits of (non-)ensuring. It is argued that currently, due to a very low level of detection of law infringement by AAs, and in consequence very low level of enforcement, the cost of ensuring the AAs are law-abiding by design exceeds the cost of not doing so.

Section Four: Towards 'Law-abiding by Design' Artificial Agents

In this section it is argued that, given the characteristics of the artificial agents, their regulation should be a mix of ex post policing (as in the 'mere tools' approach) with the ex-ante regulation. The latter would require AAs developers/users to 'code in' normative components in to the AAs code, sanctioning a lack of doing so even if AAs do not infringe the law; as well as their registration. This, however, should be facilitated by the regulators effort to create legal texts and legal databases in forms understandable by machines. The types of challenges one faces while trying to formalize the law are described (stemming both from the vagueness of legal concepts and the nature of legal reasoning), in order to provide an insight into the legal part of the puzzle for the engineers.

Conclusions

The paper concludes that, on the theoretical level, none of the available legal concepts fits well when used to refer to artificial agents, a new one is needed, and suggest a possible one. On the regulatory level, supplementing the ex-post policing with ex-ante regulation of the artificial agents' design is necessary. Questions left open, as well as suggestions for further research, especially interdisciplinary cooperation between lawyers and engineers, are enumerated.

References

Chopra, S., and White, L. F. 2011. *A Legal Theory for Autonomous Artificial Agents*. Ann Arbor: The University of Michigan Press.

Hallevy, G. 2013. *When Robots Kill: Artificial Intelligence Under Criminal Law*. Boston: Northeastern University Press.

Pagallo, U. 2013. *The Law of Robots*. Torino: Springer.

Palka, P. 2016. Redefining 'property' in the digital era: when online, do as the Romans did. EUI Department of Law Research Paper No. 2016/08.

Sartor, G. 2009a. Understanding and applying legal concepts: an inquiry on inferential meaning. In *Concepts in Law*, ed. J. C. Haage, and C. Jaap. Springer.

Sartor, G. 2009b. Cognitive automata and the law: electronic contracting and the intentionality of software agents. *Artificial Intelligence and Law* 17: 253.

Solum, L. 1991. Legal Personhood for Artificial Intelligences. *North Carolina Law Review* 70: 1231.

Sweeney, L. 2013. Discrimination in Online Ad Delivery, available at: http://dataprivacylab.org/projects/onlineads/1071-1.pdf.

Machine Ethics: Evolutionary Teachings

Luís Moniz Pereira

NOVA Laboratory for Computer Science and Informatics (NOVA LINCS),
Departamento de Informática, Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa, Portugal
lmp@fct.unl.pt

Abstract

We ponder on the teachings of human moral evolution studies for machine ethics.

Keywords: Evolutionary anthropology, utilitarianism, mutualism, contractualism, reciprocity, emergence, computational morality, Evolutionary Game Theory.

Teachings

Added dependency on cooperation makes it more competitive to cooperate well. Thus, it is advantageous to invest on shared morals in order to attract partners who will partake of mutual and balanced advantages.

This evolutionary hypothesis inspired by mutualism (Baumard 2010)—itself a form of contractualism (Ashford and Mulgan 2012)—contrasts with a number of naturalist theories of morality, which make short shrift of the importance of cognition for cooperation. For example, the theory of reciprocity, in ignoring a wider cognitive capacity to choose and attract one's partners, forbids itself from explaining evolution on the basis of a cooperation market.

Indeed, when assigning all importance to population evolutionary mechanisms, naturalist theories tend to forget the evolution of cognition in individuals. Such theories habitually start off from evolutionary mechanisms for understanding the specificity of human morals: punishment (Boyd and Richerson 1992; Sober and Wilson 1998), culture (Henrich and Boyd 2001; Sober and Wilson 1998), political alliances (Boehm 1999; Erdal et al. 1994). According to Baumard's hypothesis, morality does not emerge because humans avail themselves of new means for punishing free-riders or for recompensing cooperators, but simply because mutual help—and hence the need to find partners—becomes much more important.

In summary, it's the development of cooperation that induces the emergence of morals, and not the stabilization of morals (via punishment or culture) that promotes the development of cooperation.

Experimental results are in line with the hypothesis that the perfecting of human intuitive psychology is responsible for the emergence of morality, on the basis of an improved understanding of the mental states of others. This permits to communicate, not just to coordinate with them, and thus extend the domain cooperation, thereby leading to a disposition toward moral behaviors. For a systematic and thorough account of research into the evolutionary origins of morality, see (Krebs 2011; Bowles and Gintis 2011).

At the end of the day, one may consider three theories bearing on three different aspects of morality: the evaluation of interests for utilitarianism, the proper balance of interests for mutualism, and the discharging of obligations for the virtues principled.

A naturalistic approach to moral sense does not make the psychological level disappear to the benefit of the evolutionary one. To each its explanation level: psychology accounts for the workings of the moral sense; sociology, for the social context that activates it; and a cupola theory, for the evolution of causes that occasioned it (Sperber 1997). Moral capability is therefore a "mechanism" amongst others (Elster 1998), as are the concern for reputation, the weakness of the will, the power to reason, etc.

An approach that is at once naturalist and mutualist allows escape from these apparently opposite viewpoints: the psychological and the societal. At the level of psychological motivations, moral behavior does neither stem from egotism nor altruism. To the contrary, it aims at the mutual respect for everyone's attending interests. And, simultaneously, it obeys the logic of equity. At the evolutionary level, moral behavior is not contradictory with egotism because, in human society, it is often in our own interest to respect the interests of others. Through moral motivations, we avail ourselves of a means to reconcile the diverse individual interests. Morality vies precisely at harmonizing individual interest with the need to associate, and profit from cooperation, by adopting a logic of fairness.

The mutualist solution is not new. Contractualist philosophers have upheld it for some time. Notably, they have furnished detailed descriptions of our moral capacity (Thomson 1971; Rawls 1971). However, they never were able to explain why humans are enabled with that particular capacity: Why do our judgments seek equity? Why do we behave morally at all?

Without an explanation, the mutualist theory seems improbable: Why behave we as if an actual contract had been committed to, when in all evidence one was not?

Past and ongoing evolutionary studies, intertwining and bridging cognitive and population aspects, and both becoming supported on computational simulations, will help us find answers to that. In the process, rethinking machine ethics and its implementations.

According to (Boehm 2012), conscience and morality evolved, in the biological sense. Conscience evolved for reasons having to do with environments humans had to cope with prehistorically, and their growing ability to use group punishment to better their social and subsistence lives and create more equalized societies. His general evolutionary hypothesis is that morality began with having a conscience and that conscience evolution began with systematic but initially non-moralistic social control by groups.

This entailed punishment of individual "deviants" by bands of well-armed large-game hunters, and, like the ensuing preaching in favor of generosity, such punishment amounted to "social selection", since the social preferences of members and of groups as a whole had systematic effects on gene pools.

This punitive side of social selection adumbrates an immediate kind of "purpose", of large-brained humans actively and insightfully seeking positive social goals or avoiding social disasters arising out of conflict. No surprise the genetic consequences, even if unintended, move towards fewer tendencies for social predation and more towards social cooperation. Hence, group punishment can improve the quality of social life, and over the generations gradually shape the genotype in a similar direction.

Boehm's idea is that prehistoric humans made use of social control intensively, so that individuals who were better at inhibiting their own antisocial tendencies, by fear of punishment or by absorbing and identifying with group's rules, garnered a superior fitness. In learning to internalize rules, humankind acquired a conscience. At the beginning this stemmed from punitive social selection, having also the strong effect of suppressing free riders. A newly moralistic type of free-rider suppression helped evolve a remarkable capacity for extra-familial social generosity. That conscience gave us a primitive sense of right and wrong, which evolved the remarkable "empathy" which we are infused with today. It is a conscience that seems to be as much a Machiavellian risk calculator as a moral force that maximizes prosocial behavior, with others' interests and equity in mind, and minimizes deviance too. It is clear that "biology" and "culture" work together to render us adaptively moral.

Boehm believes the issue of selfish free riders requires further critical thought, and that selfish intimidators are a seriously neglected type of free rider. There has been too much of a single-minded focus on cheating dominating free rider theorizing. In fact, he ascertains us the more potent free riders have been alpha-type bullies, who simply take what they want. It is here his work on the evolution of hunter-gatherer egalitarianism enters, namely with its emphasis on the active and potentially quite violent policing of alpha-male social predators by their own band-level communities. Though there's a large literature on cheaters and their detection, free-rider suppression in regard to bullies has not been taken into account so far in the mathematical models that study altruism.

"For moral evolution to have been set in motion," Boehm (Boehm 2012) goes on, "more was needed than a preexisting capacity for cultural transmission. It would have helped if there were already in place a good capacity to strategize about social behavior and to calculate how to act appropriately in social situations."

In humans, the individual understanding that there exists a self in relation to others makes possible participation in moral communities. Mere self-recognition is not sufficient for a moral being with fully developed conscience, but a sense of self is a necessary first step useful in gauging the reactions of others to one's behavior and to understand their intentions. And it is especially important to realize that one can become the center of attention of a hostile group, if one's actions offend seriously its moral sensibilities. The capacity to take on the perspective of others underlies not just the ability of individuals in communities to modify their behavior and follow group imposed rules, but it also permits people acting as groups to predict and cope insightfully with the behavior of "deviants."

Social selection reduced innate dispositions to bully or cheat, and kept our conscience in place by self-inhibiting antisocial behavior. A conscience delivers us a social mirror image. A substandard conscience may generate a substandard reputation and active punishment too. A conscience supplies not just inhibitions, but serves as an early warning system that helps prudent individuals from being sanctioned.

Boehm (Boehm 2012) wraps up: "When we bring in the conscience as a highly sophisticated means of channeling behavioral tendencies so that they are expressed efficiently in terms of fitness, scenarios change radically. From within the human psyche an evolutionary conscience provided the needed self-restraint, while externally it was group sanctioning that largely took care of the dominators and cheaters. Over time, human individuals with strong free-riding tendencies—but who exercised really efficient self-control—would not have lost fitness because these predatory tendencies were so well inhibited. And if they expressed their aggression in socially acceptable ways, this in fact would have aided their fitness. That is why both free-riding genes and altruistic genes could have remained well represented and coexisting in the same gene pool."

Conclusions

For sure, we conclude, evolutionary biology and anthropology, like the cognitive sciences too (Hauser 2007; Gazzaniga 2006; Churchland 2011; Greene 2013; Tomasello 2014), have much to offer in view of rethinking machine ethics, evolutionary game theory simulations of computational morality, and functionalism to the rescue (Pereira 2016).

Acknowledgments

LMP acknowledges support from FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013.

References

Ashford, E., and Mulgan, T. 2012. Contractualism. `http://plato.stanford.edu/archives/fall2012/entries/\\contractualism/`.

Baumard, N. 2010. *Comment nous sommes devenus moraux: Une histoire naturelle du bien et du mal*. Paris: Odile Jacob.

Boehm, C. 1999. *Hierarchy in the Forest: the Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press.

Boehm, C. 2012. *Moral Origins: the Evolution of Virtue, Altruism, and Shame*. New York: Basic Books.

Bowles, S., and Gintis, H. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton: Princeton University Press.

Boyd, R., and Richerson, P. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* 13(3):171–195.

Churchland, P. 2011. *Braintrust: What Neuroscience Tells Us about Morality*. Princeton, NJ: Princeton University Press.

Elster, J. 1998. A plea for mechanisms. In *Social Mechanisms: an Analytical Approach to Social Theory*. Cambridge, NY: Cambridge University Press.

Erdal, D.; Whiten, A.; Boehm, C.; and Knauft, B. 1994. On human egalitarianism: an evolutionary product of machiavellian status escalation? *Current Anthropology* 35(2):175–183.

Gazzaniga, M. S. 2006. *The Ethical Brain: The Science of Our Moral Dilemmas*. New York: Harper Perennial.

Greene, J. 2013. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York, NY: The Penguin Press HC.

Hauser, M. D. 2007. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. London, UK: Little Brown.

Henrich, J., and Boyd, R. 2001. Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology* 208(1):78–89.

Krebs, D. L. 2011. *The Origins of Morality – An Evolutionary Account*. Oxford U. P.

Pereira, L. M. 2016. Software sans emotions but with ethical discernment. In Silva, S. G., ed., *Morality and Emotion: (Un)conscious Journey to Being*. London: Routledge.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Sober, E., and Wilson, D. 1998. *Unto Others: the Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Sperber, D. 1997. Individualisme méthodologique et cognitivisme. In *Cognition et sciences sociales*. Paris: Presses Universitaires de France.

Thomson, J. J. 1971. A defense of abortion. *Philosophy & Public Affairs* 1(1):47–66.

Tomasello, M. 2014. *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.

The predicament of decision-making in machines and humans

Mojca M. Plesničar[1] **and Danijel Skočaj**[2]

[1]Institute of Criminology at the Faculty of Law University of Ljubljana
[2]University of Ljubljana, Faculty of Computer and Information Science
mojca.plesnicar@pf.uni-lj.si, danijel.skocaj@fri.uni-lj.si

Introduction

Human decision-making is often flawed. Some reasons for that lie outside of the human reach (e.g. limited perceptual abilities and the partial observability of the environment), while others lie in the ways in which the human brain makes decisions (see e.g. Kahneman, 2011; Ariely, 2008). Machines (to use a very general term to encompass intelligent robots, software agents, etc.) may sometimes remedy the causes for the slips in human decision-making, and we already know several of such aids, but new ones are being developed as we speak. We wish to explore the ways in which machines can effectively help or even replace humans in making better decisions and the dilemmas behind such assistance.

Rule-based decision modelling

We may view the complexity of decisions as a continuum with simpler decisions on the one side and the most complex on the other with a wide range of in-between decisions. The easier task is to begin with simple decisions. What we believe is typical of them is that the issue they are addressing is well-defined, and the problem with the decision-making does not commonly stem from a human limitation in cognition, but rather in either a human limitation in perception or the environment. We can draw examples for such decisions from the realm of sports: in football the goal-line technology allows us to reliably determine whether the ball has passed the goal-line or not, something which can in certain situations be hard to see for the referees. Hawk-eye technology plays a similar role in tennis. Assisting or even replacing the human element of decision-making in such instances is quite simple and straightforward, albeit the technology behind it can be very complex. The basic approach would be rule-based: the idea is to develop a system with built in clear rules that determine what decision to make when certain preconditions are met.

There can hardly be any disagreement as to whether such a sensor system can assess whether the ball passed the line better than a human. Much doubt about their use may be mistaken – such systems typically undergo rigorous testing and are not rolled out until their fail-rate is much lower than the human fail-rate, first in parallel to human decision-makers, later on on their own. Given the high reliability and potential employability of such systems, it may seem surprising they are not used more often. However, to return to the example of sports, when deciding whether to use such a system, one needs to consider its benefits as well as its costs in terms of finance and in terms of disrupting the nature of the game. There are some voicing the complaint that, albeit improving the accuracy and even fairness of the decision, such assistance ‘kills’ the game – in other words – takes out the human element of the game, which contributes to the game being more interesting and more unpredictable.

Data-driven decision modelling

When decisions climb the complexity continuum, the issues get more tangled as well. Complex decisions mean there are no straightforward answers; there is a large variety of factors to be considered prior to making the decision, which are typically difficult to identify and elaborate, while their influences on the decision are interconnected and very difficult to measure. More and more attempts at creating machines that would assist humans in making such decisions have been rolled out in recent years. There are, for example, AI systems working at replacing juries in beauty contests,[1] algorithms replacing human editors in promoting content in social media (Fairfield & Shtein, 2014), advertising algorithms (Datta et al. 2015), and projects aimed at improving certain decisions in the criminal justice setting (e.g. releasing defendants on bail, sentencing – Bennet Moses & Chan, 2015).

The question whether it is feasible to construct a machine that could assist and improve human decision-making in such instances, is less straightforward than earlier. It seems possible, but it is much more difficult. The problem with the rule-based approach in such cases is that on the one hand, rules are not always easy to determine and often vague, and on the other hand, determining whether the rules apply in specific situations may sometimes be a very difficult task.

[1] http://beauty.ai/

Hence, a different approach is needed to address complex decisions. One very tempting option is to use machine learning, the research field that has made tremendous progress in recent years utilising the notion that algorithms can learn from and make predictions on data. Rather than explicitly program the rules to be followed in the decision making process, such an algorithm is given a number of example inputs that are used to build a model, which is later on used for making data-driven predictions or decisions. This might seem a more promising option, since it can tackle significantly more complex problems; the user does not have to describe the problem by explicitly determining the rules to follow; she only has to provide the training samples. However, two very important issues arise related to this. The set of training samples should be diverse and large enough to build a reliable model that can generalize well beyond the training samples. By increasing the complexity of the problem, the number of different factors (features, attributes) that has to be taken into account increases significantly, which, in turn, requires a huge number of training samples; a problem known as the curse of dimensionality. The other problem is that the goal of machine learning is to model the given data as well as possible while still allowing a certain level of generalisation (Kononenko & Kukar, 2007); it will therefore produce models that make decisions similar to those that occurred in the past, thus perpetuating the status quo without ever questioning it.

While building such a machine is an immensely complex task, it does seem feasible. However, this still does not determine whether we should build and utilise it. The two questions are more connected than it may appear – if we were able to address the problematic issues in advance and answer them with building in appropriate safeguards, the 'should' question might be less permeable than it appears (Kuipers, 2016).

The important issues we need to discuss tie in with both the technological backstage of machine decision-making as well as the human part. Albeit promising in many aspects, none of the previously mentioned complex AI systems seem to be working terribly well and the main problem seems to be perpetuating existing bias (beauty: white,[2] reporting: rubbish,[3] jobs: men,[4] crime: blacks[5]). One of the problems may be that explaining complex human decisions and dissecting them into the multitude of contributing factors seems to be an extremely challenging and sometimes even impossible task.

However, it could also be a consequence of machine learning with a too small or skewed data set. More importantly, it could be a consequence of the mere fact that the decisions humans make are not always the decisions we like to think we are making. In this case, the AI system is built and functioning technically correctly, however it is perpetuating the problematic decisions made by human decision-makers thus producing unwanted results on an even larger scale (O'Neil, 2016).

Conclusion

An important contribution of such AI systems is that they may pose a mirror to human decision-makers by reproducing their decisions. On the one hand, the decisions themselves can reflect the bias conducive to them thus explicitly showing the implicit reasons behind them. An even better option would be if such systems were able to explain how the bias influences final decisions. By doing so society is given an opportunity to reassess its priorities and principles upon which it makes decisions. Moreover, explainability is a highly desirable feature of AI decision-making systems for other reasons as well – for example, when humans are able to understand the causes behind the AI produced decisions, they are more likely to accept them, trust them and reflect on them.

Changing the pattern of flawed decision-making should however not be limited to only human decision-making. If we wish to build worthwhile AI systems that do not propagate such flawed decision-making, they should be able to adapt according to new evidence and evolve according to emerging new paradigms that are articulated by human decision-makers. In order to achieve that we need to build a system that is able to combine rule-based and data-driven modelling of the decision-making process. This allows for explicit rules set by humans to prevent and overrule undesired decisions that are a result of past flawed human decision-making or wrongly learnt from available data. The adaptations such systems must allow for need to be close to immediate even in the case of machine learning, despite the fact that it usually needs large sets of training samples to steer the learnt models in a new direction. However, several concepts in machine learning such as one- and zero-shot learning, knowledge-transfer etc. promise a faster process of model adaptation and should thus be considered when building complex AI decision-making systems.

Human decision-making is extremely complex and still largely unexplained. When designing AI systems we need to consider that and leave room for complementing them with new paradigms that are bound to emerge, thus enabling AI systems to serve as reliable and trustworthy aids to human decision-making.

[2] https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people?CMP=fb_gu

[3] https://www.theguardian.com/technology/2016/aug/29/facebook-fires-trending-topics-team-algorithm

[4] http://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html

[5] https://www.theguardian.com/commentisfree/2016/jun/26/algorithms-racial-bias-offenders-florida

References

Ariely. D. 2008. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. New York: HarperCollins.

Bennett Moses, L. & Chan, J. 2014. Using Big Data for Legal and Law Enforcement Decisions: Testing the New Tools. *University of New South Wales Law Journal*, 37(2), pp. 643-678.

boyd, d., & Crawford, K. 2012. Critical Questions for Big Data. *Information, Communication & Society*, 15(5), 662–679.

Datta A., Tschantz, M. T., Datta, A. 2015. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies.* 2015(1), pp. 92–112.

Fairfield, J., & Shtein, H. 2014. Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism. *Journal of Mass Media Ethics*, 29(1), 38–51.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kononenko, I., & Kukar, M. 2007. *Machine Learning and Data Mining*. Chichester, UK: Woodhead Publishing.

Kuipers, B. 2016. Toward Morality and Ethics for Robots. *Ethical and Moral Considerations in Non-Human Agents, AAAI Spring Symposium Series*.

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

No Endorsement Without Error:
Why Full Artificial Moral Agents Must Be Fallible

Frodo Podschwadek
University of Glasgow
f.podschwadek.1@research.gla.ac.uk

What does it mean for a machine or, more generally an artificial life form, to be a full moral agent? In recent years, more and more authors have aimed to define requirements for artificial moral agents (AMAs), and to identify the challenges and problems posed by various forms of artificial morality (see e.g. Allen et al. 2000, Floridi and Sanders 2004, Stahl 2004, Grau 2006, Anderson and Anderson 2007, Powers 2011, Russell et al. 2015). One definition of the central aim of machine ethics seems particularly interesting to me – the claim that "the ultimate objective for building an AMA should be to build a morally praiseworthy agent." (Allen et al. 2000).

This paper outlines what I believe to be plausible requirements for a hypothetical AMA with the potential to be morally praiseworthy. I will start by adopting the distinction between mere carriers of information and agents that hold beliefs, and the resulting conclusion that only the latter can be moral patients, i. e. entities towards whom we can have some sort of moral obligation (Anderson 2013). Contemporary moral patients falling into this category are higher animals of various kinds, but it is not hard to imagine artificial entities that at some point might also acquire the capacity to hold beliefs about the world or at least about their perceptions thereof.

Building on this distinction, I want to suggest that fully moral artificial agents resemble moral patients insofar that they hold beliefs about the world and/or about their internal representation of it. Furthermore, they would also need to have a capacity for reflection about these beliefs. This kind of reflection can be seen as the distinctive feature of moral agents in contrast to moral patients. It allows them to autonomously endorse particular beliefs which they hold as normative and therefore action-guiding (Frankfurt 1971, Korsgaard 1996). If we accept this as a definition of moral agents, it seems that only machines or other kinds of synthetic lifeforms that are capable of reflection about their beliefs and the normative status of these beliefs are genuinely ethical and autonomous agents (see e. g. Moor 2006).

It might be not clear whether and how AMAs capable of sufficiently complex meta-level reasoning could be generated, as they would also need e. g. a general phenomenal consciousness in order to have reflective capacities about their beliefs. As far as I know, the examples that we have for this kind of AMAs are mainly restricted to fiction.

If, however, AMAs that would qualify as full moral agents existed and their capacity for moral agency could be accurately described by the endorsement model given above, they would have to come with a feature that for many people would rather count as a bug in their software – they would come with the possibility of making moral errors.

It seems unlikely, although not impossible, that AMAs would act immorally due to overriding selfish motivations (see. e. g. Anderson and Anderson 2007). Instead, the potential of AMAs for moral errors would stem from the fact that ethical principles necessarily under-determine decisions (see e. g. Baier 1985, O'Neill 1987). In situations that provide hard cases of ethical questions or that are just simply novel to the agent, the AMA would have to reflect about the moral beliefs it holds and justify its course of action on the basis of this internal deliberation.

The outcome of this process could sometimes lead to immoral (or amoral) actions or consequences. Furthermore, it could lead to failure to endorse the underlying moral rules as normatively binding, either in a localized context or even on a more systemic level – resulting in moral scepticism or nihilism. Both results should be understood as moral errors, and it seems to me that full AMAs with the capacity to autonomously endorse moral principles or rules must at the same time be morally fallible AMAs in this sense.

Two questions are the consequence of this line of thought. They certainly cannot be answered in this paper, as they require more than just armchair philosophical considerations, but would also (perhaps to a higher degree) involve the expertise of engineers and maybe lawyers. Yet, at least they can be briefly outlined here.

First, if full AMAs were necessarily morally fallible, it becomes questionable whether they were desirable in the first place. One of the main incentives in constructing machines is that they are more efficient at the tasks they are supposed to do than humans. Moreover, there seems to be prima facie no conceptual problem in having machines that are able of fulfilling complex task without being capable of autonomous moral reasoning, but which could instead act according to a hard-wired set of ethical rules. These would possibly be closer to legal codes than any moral systems that are frequently discussed in moral philosophy.

It could however turn out that the capacity for fully autonomous moral deliberation occurs as a sort of emergent property together with other highly developed capacities of autonomous problem solving. It is also conceivable that intelligent systems cannot be effectively programmed in a conventional sense when they become more complex, and that it is impossible for an ethically responsible team of software engineers to make sure that every possible decision the AI makes will be bound by predefined limitations. Instead, sufficiently complex systems might have to learn by themselves and therefore be autonomous in organizing their knowledge to at least some degree. Perhaps at some point this must involve a capacity for self-reflection, or it will lead to the development of this capacity in a chain of autonomous adaptations to complex social environments. Next to the question of desirability, this scenario also leads to the second question of whether and how such a development could be prevented.

References

Allen, C.; Varner G.; and Zinser, J. 2000. Prolegomena to Any Future Artificial Agent. *Journal of Experimental and Theoretical Artifical Intelligence* 12: 251–261.

Anderson, D.L. 2013. Machine Intentionality, the Moral Status of Machines, and the Composition Problem. In *Philosophy and Theory of Artificial Intelligence*, ed. V.C. Müller, 321–333. Berlin, New York: Springer.

Anderson, M.; and Anderson, S.L. 2007. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine* 28(4): 15–26.

Baier, A. 1985. Theory and Reflective Practices. In *Postures of the Mind: Essays on Mind and Morals*: 207–227. London: Methuen.

Floridi, L.; Sanders, J.W. 2004. On the Morality of Artifical Agents. *Minds and Machines* 14*:* 349–379.

Frankfurt, H.G. 1971. Freedom of the Will and the Concept of a Person. In *The Importance of What We Care About: Philosophical Essays*: 11-25. Cambridge, New York: Cambridge University Press.

Grau, C. 2006. There Is No "I" in "Robot": Robots and Utilitarianism. *IEEE Intelligent Systems* 21(4): 52–55.

Korsgaard, C.M. 1996. *The Sources of Normativity*. Cambridge, New York: Cambridge University Press.

Moor, J.H. 2006. The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21(4): 18–21.

O'Neill, O. 1987. Abstraction, Idealization and Ideology in Ethics. *Royal Institute of Philosophy Lecture Series* 22: 55–69.

Powers, T.M. 2011. Incremental Machine Ethics. *IEEE Robotics and Automation* 18(1): 51–58.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research Priorities for Robust and Beneficial Artificial Intelligence. *AI Magazine* 36(4): 105–114.

Stahl, B.C. 2004. Information, Ethics, and Computers: The Problem of Autonomous Moral Agents. *Minds and Machines* 14: 67-83.

Property relations - Is property a person? Extending the Rights of Property Owners Through the Rights of Robotic Machines

Kathleen Richardson
Centre for Computing and Social Responsibility
De Montfort University
Kathleen.Richardson@dmu.ac.uk

Abstract

A small but vocal community in AI and robotics is calling for the recognition of the rights of machines, even suggesting that robots are slaves because they are used as instruments. I want to suggest that these scholars logic only make sense when you accept the formulations of slavery as formulated by Aristotle in the *Politics*. Advocates of slavery confuse what it means to be a person or a thing, making persons into things (instruments) for exploitation, and things (robots and AI) into persons. While formal ideas of slavery have been abolished in Europe and North America, I want to suggest that the ideas presented by Aristotle are very much present in contemporary narratives of robotics and AI. I suggest what it at work is 'property relations' which encourages humans to think of themselves as property, to re-classify property (AI and robots) as types of persons.

Rights for Machines

In the fields of AI and robotics, new narratives are arising that advocate the breakdown of distinctions between what is human and machine/person and thing. In the words of Tim Berners-Lee 'In an extreme view, the world can be seen as only as connections, nothing else' (cited in Richardson 2015, p. 1). Moreover, symmetries between humans and machines and persons and things occur because fundamentally they share no essentialist criteria, there is nothing essentially different from being a person or a thing as Donna Haraway suggested in her essay *A Cyborg Manifesto* (2006). Recently, such positions have manifested in papers that advocate for a Rights discourse to be extended to artificial agents (Gunkle, 1994). In fashionable theories that dominated some European academic departments in the late 1990s and 2000s such as actor-network theory (ANT) (Latour 2012) or transhumanism (Bostrom 2005) in different ways asserted the dissolution of distinctions between person and things. In the category of persons, I place human beings. In the category of things I include all human made artefacts. It is not possible for me to address the issue of animals in this paper, but to say that all living beings have an existence outside of property and need to addressed differently. For now we are interested in how the humanmade category of *property* is used as way of relating to humans. Robots and AI are considered highly prized humanmade artefacts, because in making these objects, techno-scientists can mobilize a framework that is distinct from magic (Musial 2016) or animistic thinking (Richardson 2016*) to justify such claims, using science fiction often as the backdrop to explore what is on the horizon technologically. What I want to suggest is philosophies which advocate the breakdown of distinctions between persons and things are revised versions of arguments for slavery. I call these revised arguments *property relations*. Property relations express the outlook of property owners, those few individuals in the global economy that significantly profit from sets of ideas that can be made use of to create new markets and new forms of exploitation. This has led to new alliances forming between academics in AI and robotics and Silicon Valley billionaires such as the *Future of Humanity Institute* (Oxford), the *Centre for the Study of Existential Risk* (Cambridge, UK) and the *Future of Life Institute* (US). In property relations, persons are things and things have the potential to become persons. Moreover, property relations is now entering a new phase of development, as human beings are encouraged to form relationships with property (robots and AI), and the person is encouraged to look upon the self (body, ideas, feelings, identity) as property. Rather than abolish slavery, these trends reveal the original idea of slavery, as advocated by Aristotle are present in contemporary narratives of robots and AI as well as critical perspectives articulated by Haraway (2006) and actor network theory (Latour 2012).

Slavery: Persons as Things

In the *Politics*, Aristotle (1992) created a formal philosophical framework for the reduction of persons to things arguing that slaves, women and children were different kinds of property of Man. In this paper, I will explore Aristotle's points about slaves as tools and demonstrate how his ideas persist today in the fields of AI and Robotics. In the *Politics*, Aristotle created a framework for AI and robotics:

1. Dissolution of distinction between persons and things

'Tools may be animate as well as inanimate…a slave is a sort of living piece of property' (cited in Richardson 2016, p. 50).

2. Non-empathetic relations

'So a slave is not only his master's slave but belongs to him tout court, while the master is his slave's master but does not belong to him' (cited in Richardson 2016, p. 51).

3. The animation of tools

'For suppose that every tool could perform its task either at our bidding or itself perceiving the need,…then master-craftsmen would have no need of servants nor masters of slaves' (cited in Richardson 2016, p.50).

Advocates of robots rights and ending robot 'slavery' suggest that these extensions of rights to artefacts come as a consequence of recognition of rights of slaves, people of colour, women and children, but on the contrary, this perspective, is inside the framework of pro-slavery. There was never a human need for slavery. There was never a need for Men to create systems of rule over another and relate to other men, women and children as property. Apologists for Aristotle suggest that slavery was a natural condition of society at the time. This is not the case. In the *Politics*, Aristotle makes one small reference to anti-slavery citizens writing 'Others say that it is contrary to nature to rule as master over slave, because the distinction between slave and free one is one of convention only, and in nature there is no difference, so that this form of rule is based on force and is therefore not just' (cited in Richardson 2016, p. 50).

People are people and things are things

Human rights discourses have developed in two distinct ways in the history of humanity. The first is to recognize that persons are not things, they are different from instruments and tools and cannot be treated as such. The second is as means to prevent rebellions (Dworkin and MacKinnon (1988). Inanimate tools (unlike like animate ones in the form of slaves) are easier to control, have fewer needs, wants and desires. An extreme form of the instrumentalisation of persons today is present in the prostitution industry. In the prostitution industry, women (the main product is women) are bought, sold, rented and traded. They are related to as things, and not persons. The buyers of these bodies for sex are allowed to relate to persons as instruments. It should be no surprise then that advocates of sex robots (Levy and Loebner 2007) make analogies between dolls, robots, general consumerist behaviors and women in prostitution. The widespread use of female and children's bodies as instruments (animate tools), and non-empathetic relational ontologies present in today's prostitution industry are carried over into sex robot narratives. I propose those who advocate on the rights of machines and encourage relationships with property (social robots, sex robots, companion robots, robot friends etc.,) should be seen as a disturbing consequence of property relations.

References

Aristotle., 1992. The Politics. Translated by T.A Sinclair, Rev and re-presented by T.J Saunders. London.: Penguin Books.

Bostrom, N., 2005. Transhumanist values. *Journal of philosophical research*, 30(Supplement), pp.3-14.

Gunkel, D.J., 2014. A vindication of the rights of machines. *Philosophy & Technology,* 27(1), pp.113-132.

Haraway, D., 2006. A cyborg manifesto: Science, technology, and socialist-feminism in the late 20th century. In *The international handbook of virtual learning environments* (pp. 117-158). Springer Netherlands.

Latour, B., 2012. *We have never been modern*. Cambridge, Mass.: Harvard University Press.

Levy, D. and Loebner, H., 2007. Robot prostitutes as alternatives to human sex workers. [Online] <Accessed 4.11.2016 http://www.roboethics.org/icra2007/contributions/LEVY%20Robot%20Prostitutes%20as%20Alternatives%20to%20Human%20Sex%20Workers.pdf>

Richardson, K., 2015. *An Anthropology of Robots and AI: Annihilation Anxiety and Machines* (Vol. 20). New York.: Routledge.

Richardson, K., 2016*. Technological Animism: The Uncanny Personhood of Humanoid Machines. *Social Analysis, 60*(1), pp.110-128.

Richardson, K., 2016. Sex Robot Matters: Slavery, the Prostituted, and the Rights of Machines. *IEEE Technology and Society Magazine,* 35(2), pp.46-53.

Dworkin, A. and MacKinnon, C.A., 1988. *Pornography and civil rights: A new day for women's equality*. Organizing Against Pornography.

Musial, M., 2016. Loving Dolls and Robots: From Freedom to Objectification, From Solipsism to Autism? In John T. Grider and Dionne van Reenen (eds.) *Exploring Erotic Encounters: The Inescapable Entanglement of Tradition, Transcendence and Transgression*. Oxford.: Inter-Disciplinary Press

Autonomous weapons systems on a post-heroic battlefield. Ethical and legal dilemmas

Błażej Sajduk, PhD
Jagiellonian University
blazej.sajduk@uj.edu.pl

Abstract

The presentation will embed the issue of artificial agents in a broader context, I will indicate ethical and legal problems of deploying Lethal Autonomous Weapons Systems (LAWS) on a battlefield. The International Relations and war theory will be also taken into consideration.

The appearance of the Unmanned Combat Aerial Vehicles (UCAV) was a part of the process of withdrawing of armies of highly economically developed countries from the universe shared with the opponent. The logical consequence and at the same time a revolutionary change crowning this process would be including fully autonomous weapon systems in warfare. Using them on the battlefield will be, from the point of view of criminal and civil law, a challenge which is especially urgent in the context of assigning responsibility. As regards international humanitarian law, the problem will lie in ensuring respect for the principles of distinguishing between combatants and civilians as well as proportionality. The emergence of LAWS in military service will be the last stage in the centuries-old process of moving away dangers from one's own forces, and will be tantamount to dehumanising wars and soldiers fighting them. Together with the people, also constitutive elements of human nature, including, among others, responsibility, fear and bravery will also be removed from warfare. This may contribute to instrumentalising war, which will cease to have its human and, thus, ethical dimension, and will begin to be considered solely in terms of effectiveness and procedural efficiency. This could also mean the end of the world of values in which we have lived since the dawn of history, and the emergence of a completely new order, which is difficult to imagine today.

The dehumanization of the way in which the so-called "Western" countries wage wars is currently a widely discussed matter in academic spheres. This phenomenon consists largely of two strongly interrelated processes. The first one was described in the mid nineties by Edward Luttwak as post-heroic warfare or de-heroisation of battlefield. The second aspect is the growing autonomy of the military equipment employed. Both are indirectly the result of age-old trend shaping the direction of the development of weapons, aiming at reducing risks associated with the conduct of military operations by increasing the distance from which the opponent is attacked. Today, thanks to the development of new technologies, one may observe the next stage of the process. In contrast to earlier phases, now we may be witnessing a change of a qualitative nature. The increase in distance of the fighting parties from the war theatre has grown so much that now one speaks of the phenomena described by Christopher Coker as "disconnecting" soldiers from the battlefield. Robert Sparrow aptly noted that we find ourselves in the process of not only moving away the risk from those conducting military operations, but also removing risk from such operations, which makes it increasingly more difficult to convince the public about the value of courage and sacrificing one's life in war. From this perspective, the idea of bestowing armed machines with autonomy is a consequence and the culmination of the process of moving away from the opponent, leading to complete dehumanization of the battlefield. Of course, in this context, the fundamental objections are raised by the possibility transferring decisions regarding human to life to LAWS.

References

Arkin R.C. 07 2013. issue 137. *Lethal Autonomous Systems and the Plight of the Non-combatant*, "AISB Quarterly". http://www.unog.ch/80256EDD006B8954/%28httpAssets%29/54B1B7A616EA1D10C1257CCC00478A59/%24file/Article_Arkin_LAWS.pdf.

Arkin R.C. 03 2014. *Speaker's Summary. Ethical Restraint Of Lethal Autonomous Robotic Systems: Requirements, Research, And Implications* [in:] *Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects*. Geneva. https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf.

Asaro P. 03 2014. *Ethical Issues Raised By Autonomous Weapon Systems* [in:] *Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects*. Geneva. https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf.

International Committee of the Red Cross. 2014. *Autonomous Weapon Systems. Technical, Military, Legal and Humanitarian Aspects.* Geneva.

https://www.icrc.org/en/download/file/1707/4221-002-autonomous-weapons-systems-full-report.pdf.

Beck U. 2002. Społeczeństwo *ryzyka. W drodze do innej nowoczesności [Society of risk. On the way to different modernity]*. Wydawnictwo Scholar. Warszawa.

Chamayou G. 2015. *A Theory of the Drone*. The New Press. New York-London.

Coker C. 2008. *Ethics and War in the 21st Century*, Routledge, London – New York.

Coker C. 2001. *Humane Warfare*. Routledge. London–New York.

Coker C. 2014. *Man at War. What Fiction Tells Us about Conflict, from the Illiad to Catch-22*, Oxford University Press. New York.

Coker C. 2007. *The Warrior Ethos. Military Culture and the War on Terror*. Routledge. London-New York.

Coker C. 2009. *War in an age of risk*. Polity Press. Cambridge.

Coker C. 2013. *Warrior Geeks. How 21st-Century Technology is Changing the Way We Fight and Think About War*. Columbia University Press. New York.

Department of Defense Manual No 1348.33 vol. 3. 2013. http://www.dtic.mil/whs/directives/corres/pdf/134833vol2.pdf.

Herbach J.D. 2012. *Into the Caves of Steel: Precaution, Cognition and Robotic Weapon Systems Under the International Law of Armed Conflict*, "Amsterdam Law Forum". http://conflictandsecuritylaw.org/web_documents/277-1619-1-pb.pdf.

Heyns C. 2013. *Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions*. https://www.legal-tools.org/uploads/tx_ltpdb/A-HRC-23-47_en.pdf.

Kant I. 2013. *Uzasadnienie metafizyki moralności* [*Groundwork of the Metaphysics of Morals*]. Wydawnictwo Marek Derewiecki. Kęty.

Lee P. 2012. *Remoteless, Risk and Aircrew Ethos*. http://www.airpowerstudies.co.uk/sitebuildercontent/sitebuilderfiles/aprvol15no1.pdf.

Lin P., Abney K., Bekey G. A. eds. 2012. *Robot Ethics The Ethical and Social Implications of Robotics*. London.

Loosing Humanity. 2012. https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots.

Luttwak E. 1995. *Toward Post-Heroic Warfare*. 'Foreign Affairs'.

Luttwak, E. 1999. *"Post-Heroic Warfare" and Its Implications*. http://www.nids.go.jp/english/event/symposium/pdf/1999/sympo_e1999_5.pdf .

Matthias A. 2004. *The responsibility gap: Ascribing responsibility for the actions of learning automata*, "Ethics and Information Technology".

Memorandum: Distinguished Warfare Medal. 2013. http://www.defense.gov/news/DistinguishedWarfareMedalMemo.pdf.

Münckler H. 2004. *Wojny naszych czasów* [*Wars of our times*]. Wydawnictwo WAM. Kraków.

Ossowska M. 2000. *Etos rycerski i jego odmiany* [*Chivalric ethos and its variations*]. Wydawnictwo Naukowe PWN. Warszawa.

Scharre P. 2016. *Autonomous Weapons and Operational* Risk. Centre for a New America Security. http://www.cnas.org/sites/default/files/publications-pdf/CNAS_Autonomous-weapons-operational-risk.pdf.

Shaking the Foundations. 2014. https://www.hrw.org/report/2014/05/12/shaking-foundations/human-rights-implications-killer-robots.

Lucas G. eds. 2015. *Routledge Handbook of Military Ethics* . Routledge, London-New York.

Sparrow R. 2007. *Killer robots*. "Journal of Applied Philosophy". http://profiles.arts.monash.edu.au/wp-content/arts-files/rob-sparrow/KillerrobotsForWeb.pdf.

Toffler A. 1997. *Trzecia fala* [*The Third Wave*]. Polski Instytut Wydawniczy. Poznań.

Wagner M. 2014 vol. 47. *The Dehumanization of International Humanitarian Law: Legal, Ethical, and Political Implications of Autonomous Weapon Systems*. "Vanderbilt Journal of Transnational Law". https://www.law.upenn.edu/live/files/4003-20141120---wagner-markus-dehumanizationpdf.

Smit I. I., Lasker L., Wallach W. 2005 Vol. III. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*. http://www.realtechsupport.org/UB/WBR/texts/Wallach_ArtificialMorality_2005.pdf.

Zakrzewski L. S. 2004. *Etos rycerski w dawnej i współczesnej wojnie* [*Chivalric ethos in past and contemporary wars*]. Trio. Warszawa.

Machine Ethics: Theory vs Practice

Georgi Stojanov
The American University of Paris, France
gstojanov@aup.edu

Abstract

This short paper sketches a bigger project which explores the impact that the research in machine ethics had on the practice of using and cohabitating with autonomous systems, as well as concerns regarding their design. The overall message is that despite the booming of the field (at least in the last 15 years) and at least 30 years of research, we are far from coming up with intelligent machines equipped with a moral compass. The bigger project mentioned above will try to identify the main reasons why the research in machine ethics seems to have had none or minimal impact on the actual autonomous systems, as well as suggest ways for moving on.

Theory vs Practice

"Mercedes, Google, Volvo To Accept Liability When Their Autonomous Cars Screw Up", (Ballaban, 2015). This is a title which, among many others along the similar lines, appeared in the press about a year ago, in Fall 2015. With a simple (?) declaration these industry leaders, in essence, solved in a pragmatic way, the "liability issue" so often discussed in the machine ethics (and in particular, autonomous cars) research. In a way, this summarizes the main message of this short paper which, in a nutshell, is that the solutions to most of the issues raised in the literature will be solved in an evolutionary way, based on precedencies, which will follow the inevitable penetration of autonomous machines in every sphere of our lives. This democratization of technology will face us with unanticipated situations and the legislation will slowly follow. An example of such un anticipated problem was that silent electric cars represent a serious problem for inattentive pedestrians. Just recently (November 2016) the US National Highway Traffic Safety Administration (NHTSA) has announced a rulw which will require electric vehicles to make noise at low speeds. The rule is aimed to prevent accidents by warning pedestrians, especially those with visual impairments (http://www.rt.com/366998-electric-cars-noise-ruling)

The appearance of self-driving cars was by no account sudden. It was an evolutionary process where humans began to gradually leave more and more functions to the car itself (from ignition, to switching gears, to limiting the speed below the legal threshold, to locking and self-parking…). Brian Ladd in his book "Autophobia" (2008) gives a fascinating account of the history of cars (inventions, public acceptance, legislations…).

Now, when we are at the last frontier, giving the car the power for full navigation control, even in urban environments, there is definitely a qualitative change. We are understandably concerned whether such a car will always (or, at least, in most cases) make the "right" decision in every specific situation. Many, even quite recent articles (e.g. Goodal, 2014) immediately start with the liability and ethical decisions issues:

> "The first problem is liability, as it is currently unclear who would be at fault if a vehicle crashed while self-driving. The second problem is the ability of an automated vehicle to make ethically-complex decisions when driving, particularly prior to a crash." (Goodal, 2014)

Throughout the paper though, the liability issue is barely mentioned, and instead the author tries to give answers to possible objections about the utility of research in machine ethics (in the case of self-driving cars). In the continuation, he gives potential framework for continuing the research in the domain: from Kantian ethics, to utilitarianism, to Smithianism, commenting on potential issues with each of those. Towards the end, the author also gives a few examples of rare cases of actually implemented "ethics modules".

I want to argue that we have been gradually outsourcing decision making (even decisions with potential high risk ethical consequences) to technology. Much along the lines Andy Clark argues in "Natural Born Cyborgs" (1999), we, human beings (or better: the human minds) are opportunistic. Take the humble alarm clock which can be very simple or quite complex (e.g. an app on a smart phone) which has certain autonomy though it's restricted in its possible actions: if functioning properly it will wake you up us at the desired time. Now, we can certainly imagine a scenario where something goes wrong and the alarm doesn't ring in the morning. Depending on who you are and what you were supposed to do that morning, the damage can be considerable. Is there a case to be made here about the liability of the alarm or the company who made it? Moving on towards a slightly more complex but equally ubiquitous gadget: GPS or the car navigation systems. We let ourselves be guided by these routinely without giving it much thought. Certainly, sometimes we can state preferences for certain routes but this does not prevent the GPS to still suggest alternative routes. It has definitely certain autonomy. Many factors go into its decisions. The software is usually proprietary and a black box for the user. Given the choice of two alternate highway routes from A to B, how can we be sure that the GPS navigator manufacturer has not struck a deal with a company which has gas stations, hotels, or restaurants on along the highway of one of the alternative routes, that a route which is preferable for the said company would be the first suggestion? This and related questions may appear meta-ethical but with the increasing complexity of the technology the frontiers between ethics and meta—ethics will become even more blurred.

Elsewhere (Kinne&Stojanov, 2016) we have discussed some of the issues arising with the appearance of the autonomous lethal weapons. Among our (pragmatic) suggestions there was the **transparency:** if ever machines become better in distinguishing legitimate targets from noncombatants, whatever "ethical module" guides these weapons, it is an imperative that the module should be absolutely transparent for potential investigations which may occur afterwards.

I want to mention also here the sheer intractability of the problem of making the "right ethical decision" by comparing it to contextual object/scene recognition in AI. We are yet far from perfect (or even acceptable) object recognition: a simple apple can be described as "apple" or "Granny Smith" or "fruit" etc. depending on situational cues and humans excel in this. Machines are not even close. The existence of philosophical moral and ethical theories may give a false initial hope that it might be relatively easy to translate them into working guiding "ethics modules" but the practice has shown otherwise.

Moving on to another domain: using robots in elderly (or otherwise incapacitated) humans. In their long article, Sharkey (2010) puts forward the main ethical concerns in the domain:

> "There are two main ethical concerns about the use of assistive robot care for the elderly and its effects on their welfare – first that it might reduce the amount of human contact that the elderly have, and second that if used insensitively, it could increase senior citizens' feeling of objectification and a lack of control over their lives." (Sharkey, 2010)

These are obviously legitimate concerns, but one can argue that introducing robots in this milieu is far from being the major cause for these problems. On the contrary, taking the example of the PARO robot designed to increase the quality of life among elderly, appears to serve the purpose. Rigorous scientific evaluation of the potential benefits have started to appear and below is a quote from a pilot study, done as a preparation for a much larger assessment project (Yu et al. 2015):

> "The analysis showed that the frequency of neutral expressions during the six 30-minute sessions was high (mean observed frequency of the six sessions was 27 minutes out of 30 minutes), followed by smile (13 minutes out of 30 minutes), and laugh (9 minutes out of 30 minutes). Furthermore, all subjects gently stroked or held PARO during the interaction, and talked directly with PARO in a dyadic relation as if it was a real living pet. In addition, there was a positive trend in depressive symptoms, as evaluated by CSDD (P=.03) and a falling trend in caregiver burden, as evaluated by ZBI (P=.02) immediately following the PARO therapy."

There is an estimate of about 1300 sold in Japan from 2005 to 2010 only to individuals as well as elderly care units, as well as in Europe and the USA (parorobots.com).

Apart from these robots like PARO, designed with a special purpose in mind, there are other, general-purpose robots that are appearing on the market. Pepper, produced by Aldebaran/SoftBank Robotics, is a humanoid robot (with its height of about 1.2 m and weight of about 30 kg.) but with intentional comic--book cute looks, appeared on the market in June 2015 and about 10,000 robots were sold (Pandey, personal communication). Pepper is advertised as being able to "read human emotions" but there are no specific tasks it is intended to be used. Currently, there are only several hundreds of "apps" that users can download. Only the future will tell what kind of usage will people find. What is exciting is the fact that, probably for the first time in history, we have so many homes equipped with sophisticated "home-robots" and our community should look forward for case studies as well as the above mentioned emergent phenomena which will probably become the precedents around which machine ethics, morality, and legislation will be founded.

In summary, the technological advances in autonomous machines are inevitably advancing. What can be done, will be done. The solution of the emerging ethical and moral issues will come gradually as we will be adopting them and will be encountering new and unanticipated situations. As in most cases, the legislation will be trying to catch up with the new realities and, most likely, be based on precedents and not on well-crafted monolithic set of generic rules.

.

References

Noah J. Goodall, N. J., 2014. “Machine Ethics and Automated Vehicles”, in G. Meyer and S. Beiker (eds.), *Road Vehicle Automation,* Springer, 2014, pp. 93-102. Available at http://dx.doi.org/10.1007/978- 3-319-05990-7_9

Ballaban, M., 2015. “Mercedes, Google, Volvo To Accept Liability When Their Autonomous Cars Screw Up”, http://jalopnik.com/mercedes-google-volvo-to-accept-liability-when-their-1735170893

Ladd, B. 2008. Autophobia: *Love and Hate in the Automotive Age*, The University of Chicago Press.

Amigoni, F., and V. Schiaffonati, 2005. “Machine Ethics and Human Ethics: A Critical View” in Technical Report FS-05-06, AAAI Fall Symposium on Machine Ethics 2005, AAAI Press

Sharkey, A., 2010. “Granny and the robots: Ethical issues in robot care for the elderly”, in Ethics and Information Technology, · March 2010

Yu, R., Hui, E., Lee, J., Poon, D., Ng, A., Sit, K., Woo, J., 2015. “Use of a Therapeutic, Socially Assistive Pet Robot (PARO) in Improving Mood and Stimulating Social Interaction and Communication for People With Dementia: Study Protocol for a Randomized Controlled Trial”, JMIR Research Protocols, 4(2), e45. http://doi.org/10.2196/resprot.4189

Moral Code: Programming the Ethical Robot

Sean Welsh
Department of Philosophy, University of Canterbury, Christchurch, New Zealand
sean.welsh@pg.canterbury.ac.nz

Extended Abstract

The aim of this paper is to show how a robot can pass reasonable person tests. Passing such tests will demonstrate measurable levels of moral competence. A robot does not need "full virtue" - the ability to do the right thing for the right reasons with the right feelings (Hursthouse 1999) - to pass such tests. It only requires the ability to do the right thing for the right reasons with no feelings.

Passing a reasonable person test will not make a machine a person, merely a machine than can pass one reasonable person test. Unlike the Turing test, reasonable person tests are particular not general. A robot might demonstrate a high degree of moral competence in a bar but lack programming for medical or military contexts and thus demonstrate no competence in a hospital or battlespace. Designing a robot to pass a Turing test or a moral Turing test is not the goal. Rather the goal is to design a robot capable of passing many reasonable person tests. These are objective tests of moral competence in specific domains.

A method of test driven development as applied to machine ethics is proposed. The method requires the definition of a set of tests with machine-readable input and machine-actionable output. These tests take the form of moral dilemmas. To pass the tests the "moral code" developed has to select the "right" output given the input.

The method is agnostic on moral theory, knowledge representation and reasoning. Anything goes on the whiteboard. However, once you start to implement, ethical and technical commitments must be made.

To pass a reasonable person test the embodied robot first needs to pass symbol grounding tests (tests on individual symbols), then specific norm tests (tests on individual rules) and then finally reasonable person tests (tests involving multiple clashing rules).

The ability to pass "reasonable person" tests requires the agent have the ability to determine the "spirit of the law" as well as the "letter of the law" (Lucas 1963). In AI terms, the "spirit of the law" is the ability to infer a policy rule on the fly that is consistent with the goals sought by the "letter of the law". What is "reasonable" can override the letter of the law. For example, an agent can engage in prohibited acts such as committing trespass and doing wilful damage (e.g. smashing a window or breaking a door) to rescue people from a burning house. The lesser goals of privacy and protection of property sought by the torts of trespass and wilful damage can be overridden by the greater goal of preserving life. A clear basis for prioritization (deciding what matters more) is critical to passing reasonable person tests.

Several scenarios are presented. *Speeding Camera* and *Bar Robot* are specific norm tests. *Postal Rescue* is a culturally invariant and morally obvious reasonable person test. *Amusement Ride* is culturally invariant and arguable (i.e. not obvious) reasonable person test. Graph-based knowledge representations (Chein and Mugnier 2008) and first order logic are used to formalize the problems and pass the tests.

Several well-known trolley problems are formalized: namely *Cave*, *Hospital*, *Switch* and *Footbridge*. These tests are minimally variant by culture but arguable. Lexical orderings (Rawls 1972) based on needs theory (Reader 2007) and the ethics of risk (Hansson 2014) are introduced to solve these problems.

Having formalized several culturally invariant tests, we set about formalizing tests where culture makes a difference to right and wrong. We formalize both versions of "right" and examine the differences in formalizations. Similarly we formalize tests where an agent is morally defective and arrives at "wrong" decisions. Variance and failure result from divergent argument graphs.

Following Pigden (1989) the view that moral reasoning requires a superset of FOL is rejected. The reasoning is held to first order logic (FOL). Deontic concepts are expressed with binary predicates (as relations between agents and acts). Agents, patients, acts and valued goals are explicitly represented. Following a suggestion of Castañeda (1981) imperatives are explicitly represented as well. This avoids many "paradoxes" of deontic logic.

The expressivity necessary to solve the test problems is achieved by means of a complex knowledge representation (a graph database) rather than complex reasoning (a deontic logic that is a superset of FOL). The graph-based knowledge representation defines state-act-state transition relations (for planning), casual relations (for determining what is "reasonably foreseeable"), classification relations (for class membership inferences) and evaluation relations (for estimating the "moral force" or the "weight" of a reason).

The output of the test driven development method of machine ethics is a hybrid moral theory that employs elements of deontology, needs theory, utilitarianism, virtue ethics and contractualism to pass the tests.

References

Castañeda, H.-N. (1981). The Paradoxes of Deontic Logic: The Simplest Solution to All of Them in One Fell Swoop. New Studies in Deontic Logic. R. Hilpinen. Dordrecht, D. Reidel Publishing Company**:** 37-86.

Chein, M. and M.-L. Mugnier (2008). Graph-based knowledge representation: computational foundations of conceptual graphs, Springer Science & Business Media.

Hansson, S. O. (2014). The Ethics of Risk: Ethical Analysis in an Uncertain World. Basingstoke: New York, Palgrave Macmillan.

Hursthouse, R. (1999). On virtue ethics, Oxford University Press.

Lucas, J. R. (1963). "The Philosophy of the Reasonable Man." The Philosophical Quarterly (1950-) **13**(51): 97-106.

Pigden, C. R. (1989). "Logic and the autonomy of ethics." Australasian Journal of Philosophy **67**(2): 127-151.

Rawls, J. (1972). A Theory of Justice. Oxford, Clarendon Press.

Reader, S. (2007). Needs and Moral Necessity. London; New York, Routledge.

Lack of effort or lack of ability? Robot failures and human perception of agency and responsibility

Sophie van der Woerdt[a] **Willem F.G. Haselager**[b]

[a]Dpt. of Psychology, [b]Dpt. of Artificial Intelligence, Radboud University,
Comeniuslaan 4, 6525 HP, Nijmegen

Abstract

Research on human interaction has shown that attributing agency to another agent has substantial consequences for the way we perceive and evaluate its actions. Specifically, considering an agent's actions related to either effort or ability can have important consequences for the attribution of responsibility. This study indicates that participants' interpretation of a robot failure in terms of effort –as opposed to ability– significantly increases their attribution of agency and –to some extent– moral responsibility to the robot. However a robot displaying lack of effort does not lead to the level of affective and behavioural reactions of participants normally found in reactions to other human agents.

Introduction

Currently, much debate is devoted to the question of how we should deal with harm caused by robots (Asaro 2013; Singer 2011). Research on anthropomorphism (Duffy 2003; Złotowski , Strasser & Bartneck 2014), blame (Moon & Nass 1998; Serenko 2007; Kim & Hinds 2006; You, Nie, Suh & Sundar 2011; Koay, Syrdal, Walters & Dautenhahn 2009; Vilaza, Haselager, Campos, & Vuurpijl 2014; Malle, Scheutz, Arnold, Voiklis, & Cusimano 2015; Malle, Scheutz, Forlizzi, & Voiklis 2016) and examples of media and pop culture speaking of 'robot laws' (Clarke 1994) underline the possibility of humans ─perhaps inappropriately─ attributing moral responsibility to automated systems. Although legal solutions have been proposed for dealing with such conflicts (Asaro 2013), in daily life this may still have undesired implications. Owners and developers of robots may (unknowingly) distance themselves from potential harms caused by their robots (Coleman 2004), causing responsibility to become diffused. Therefore, it is relevant to find out what factors contribute to the attribution of agency and responsibility in robots.

Extensive work on attributional processes in human interaction reveals that the perception of an agent's effort and abilities are central determinants in the attribution of agency and moral responsibility (Weiner 1995). This, in turn, is strongly related to fundamental affective and behavioural reactions such as sympathy, rejection, altruism and aggression (Weiner 1995; Rudolph, Roesch, Greitemeyer, & Weiner 2004). Yet, with regard to human robot interaction (HRI), little is known about the attribution of agency and moral responsibility.

In this study, we applied Weiner's *Theory of Social Conduct* (Weiner 1995) to HRI by showing participants videos of robots (Aldebaran's NAO) failing tasks in ways that could be interpreted as due to either *lack of ability* (LA-condition; e.g. dropping an object) or *lack of effort* (LE-condition; e.g. throwing away an object, fig. 1). We expected that a display of *lack of effort* would incite the illusion of a robot having *agency* over its actions. In addition, we expected that a robot's *lack of effort* would have little effect on the attribution of moral *responsibility* to the robot, compared to a display of *lack of ability*.

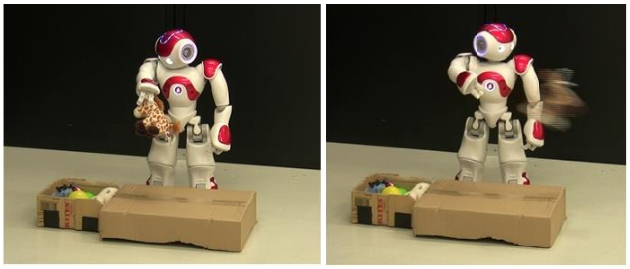

Fig. 1: sample frames of (a) a robot displaying the intention of putting a toy in a box, (b) the robot throwing the toy away instead (LE-condition).

Method

In an online survey, sixty-three participants (M_{Age} = 25,5, *SD* = 9,7; drawn from a university population) were shown a video of about 30-60 seconds portraying a situation in which a NAO robot was shown failing a task either due to *lack of ability* or *lack of effort*. Seven of such scenarios were presented[1]. After each video, participants were asked to fill in a questionnaire containing scales of *agency* (five questions about the robot's control over the situation and its ability to make its own decisions), and *responsibility* (twelve questions on attributed blame and kindness, affective and behavioural reactions). Additionally, scales were included measuring the participant's estimate of the robot's *experience* (e.g. having beliefs, desires, intentions, emotions), *predictability*, *propensity to do damage*, *trustworthiness* and *nonanthropomorphic features* (e.g. strength, efficiency, usefulness)[2].

For analysis, mean scores of each scale (range 1-5) were calculated and transposed to Z-scores. Since reliability and goodness-of-fit for the scale of *responsibility* was questionable, items of this scale were analyzed separately. In order to answer our main questions, a GLM multivariate analysis was performed with the composite means of *agency*, *experience*, *predictability*, *propensity to do damage,* and each item related to *responsibility* as dependent variables. *Condition* (LA/LE) was indicated as between-subject factor.

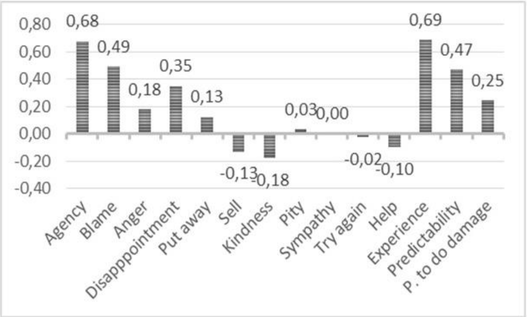

Fig. 2: Difference scores of means within the *lack of ability* and *lack of effort* (LA - LE) conditions.

Results

According to what was expected, participants attributed more agency to a NAO robot after seeing videos in which it displayed *lack of effort* (*M* = 2.80, *SD* = 0.82) compared to videos in which it displayed *lack of ability* (*M* = 2.12, *SD* = 0.61). Univariate tests expressed significant and large effects for the composite scores of *agency* (*F*(1,61) = 13.601, *p* = .000, eta^2 = .182), *experience* (*F*(1,61) = 12.235, *p* = .001, eta^2 = .168), and *predictability* (*F*(1,61) = 14.040, *p* = .000, eta^2 = .187). The results for the items of *responsibility* were mixed. While univariate tests for *blame* and *disappointment* revealed significant, medium effects (respectively: *F*(1, 61) = 5.757, *p* = .019, eta^2 = .086; *F*(1, 61) = 9.704, *p* = .003, eta^2 = .137), effects for the items *anger*, *put away, sell, kindness, pity, sympathy, help* and *try again* were not significant.

Conclusion

Similar to findings related to human interaction, the results of our study reveal that, in case of robots displaying behaviour that can be interpreted as lack of effort, humans tend to explain robotic behaviour by attributing agency. In case of failure, a robot displaying lack of effort ─essentially refraining from 'trying'─ may lead to blame and disappointment. However, it does not necessarily lead to negative affective and behavioural reactions such as anger, or wanting to shut the robot off and put it away. Results like these emphasize that we should be aware of potential diffusion of human responsibility when (advanced) robots create the impression that they are agents in the sense of actually controlling and intending their own actions. Our results also suggest that –in case of NAO robots─ failure, or even reluctance for doing tasks is received well, illustrating a promisingly positive view on robots.

References

Asaro, P. 2013. A body to kick, but still no soul to damn: Legal perspectives on robotics. In *Robot Ethics: The Ethical and Social Implications of Robotics,* ed. Lin, K. Abney, and G. Bekey, 169-186. Cambridge: MIT Press.

Clarke, R. 1994. Asimov's Laws for Robotics: Implications for Information Technology. Parts 1 and 2. *Computer,* part 1: 26(12): 53-61, 1993 and part 2: 27(1): 57-65), 1994.

Coleman, K.W. 2004. Computing and Moral Responsibility. In *The Stanford Encyclopedia of Philosophy (Fall 2006 Edition)*, ed. Zalta, E.N. Stanford: The Metaphysics Research Lab.

Duffy, B.R. 2003. Anthropomorphism and the social robot. *Robotics and Autonomous Systems,* 42(3-4): 177–190.

Kim, T.; and Hinds, P. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In Proceedings of ROMAN 06: The 15th IEEE International Symposium on Robot and Human Interactive Communication, 80-85. IEEE.

Koay, K.L.; Syrdal, D.S.; Walters, M.L.; and Dautenhahn. K. 2009. Five weeks in the robot house – exploratory human-robot interaction trials in a domestic setting. In Proceedings of the 2009 Second International Conferences on Advances in Computer-Human Interactions, 219-226. IEEE.

[1] Videos and complete survey can be found online: https://www.youtube.com/playlist?list=PLSQsUzV48QtG__YPY6kVcgCM8-YOcNqja; https://eu.qualtrics.com/jfe/preview/SV_6y4TuTii0CFnpch

[2] In this abstract we chose to focus on our main questions only. Therefore, additional analyses and results regarding these variables will not be discussed.

Malle, B.F.; Scheutz, M.; Arnold, T.; Voiklis, J.; and Cusimano, C. 2015. Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, 117-124. New York: ACM.

Malle, B.F.; Scheutz, M.; Forlizzi, J.; and Voiklis, J.;. 2016. Which Robot Am I Thinking about? The Impact of Action and Appearance on People's Evaluations of a Moral Robot. In The Eleventh Annual Meeting of the IEEE Conference on Human-Robot Interaction, 125-132. IEEE.

Moon, Y.; and Nass, C. 1998. Are computers scapegoats? Attributions of responsibility in human computer interaction. *International Journal of Human-Computer Interaction*, 49(1): 79–94.

Rudolph, U.; Roesch, S.C.; Greitemeyer, T.; and Weiner. B. 2004. A meta-analytic review of help giving and aggression from an attributional perspective. *Cognition and Emotion*, 18(6): 815–848.

Serenko, A. 2007. Are interface agents scapegoats? Attributions of responsibility in human-agent interaction. *Interacting With Computers,* 19(2): 293–303.

Singer, P.W. 2011. Military robotics and ethics: A world of killer apps. *Nature,* 477(7365): 399–401.

Vilaza, G.N.; Haselager, W.F.G.; Campos, A.M.C; and Vuurpijl, L. 2014. Using games to investigate sense of agency and attribution of responsibility. In Proceedings of the 2014 SBGames, 393-399.

Weiner. B. 1995. *Judgments of responsibility: A foundation for a theory of social conduct*. New York/London: Guilford Press.

You, S.; Nie, J.; Suh, K.; and Sundar, S. 2011. When the robot criticizes you: Self-serving bias in human-robot interaction. In Proceedings of the 6th International Conference on Human Robot Interaction, 295-296. IEEE.

Złotowski, J.; Strasser, E.; and Bartneck, C. 2014. Dimensions of anthropomorphism: from humanness to humanlikeness. In Proceedings of the 2014 ACM/IEEE international conference on human–robot interaction, 66-73. New York: ACM.

Ethical Concern of Brain–computer Interfaces

Katerina Zdravkova
University “Sts Cyril and Methodius”, Faculty of Computer Science and Engineering, Skopje, Macedonia
katerina.zdravkova@finki.ukim.mk

Abstract

Brain-computer interfaces enable a communication between neuromuscular system and peripheral nerves and muscles. They are implemented to establish or improve the missing or lost functions of the disabled, or to enhance the functions of healthy people. Brain implants interfere the natural brain-human interaction, raising various ethical questions. This paper presents some of them, intending to increase the awareness of their prospective erroneous or dishonest use.

Introduction

In the dawn of the 21st century, Bill Joy posed the question “Why the future doesn’t need us?” (Joy, 2000). He was deeply concerned that the most powerful new technologies might cause knowledge-enabled mass destruction. Due to the obvious exponential growth of technology, singularity is rapidly approaching (Vinge, 1993). It can seriously endanger its creators, the humans, particularly in the moment when they transcend biology (Kurzweil, 2005). Kurzweil’s Law of Accelerating Returns (Kurzweil, 2001) supports his claim that the evolution steadily reaches the fifth epoch, the moment when human technology starts merging human intelligence (Kurzweil, 2005). The law confirms Clark’s earlier conviction that we are already natural born cyborgs (Clark, 2001). Recent advances of brain-computer interfaces prove Clark’s assertions that we are steadily becoming human-technology symbionts, and Kurzweil’s estimation that the moment of uniting the technology with human intelligence is near (Hassanien, 2015).

Brain-computer (BCI) or brain-machine interface (BMI) is a technology that enables communication with the brain’s neuromuscular output channels of peripheral nerves and muscles (Wolpaw, 2002). It neither depends on the neuromuscular output channels, nor interferes brain’s electro-magnetic activities (Van Erp, 2012).

BCI was initially developed for medical purposes: to monitor and record electrical activity of the brain; to measure the magnetic fields generated by neuronal activity of the brain; to enable brain-control of prosthetic extremity devices; to stimulate the brain in paralysis and in Parkinson’s disease (Palferman, 2016).

Beyond these applications, many assistive technologies were invented to support communication with the visually or hearing impaired (Rupp, 2014). Furthermore, brain-machine interfaces can be used by healthy people, for example, to enhance brain cognitive functions (Lebedev, 2014). The ultimate intention of these inventions is to increase life expectancy and vastly improve the quality of life. Will some of the recent research programs, such as European Commission’s Brain research (EC, 2007-2012) and DARPA’s Brain Initiative (DARPA, 2013) collapse all the previous results like a house of cards? Potential threats of interacting with the brain have not been recognized by the legislation systems yet, increasing the risk of a total devastation of human lives whenever the outcomes of BCI research are not ethically used.

The goal of this paper is to raise the awareness of the potential misuse of BCI. It initiates a discussion about the most threatening aspects of brain implants and offers arguments that challenge the ultimate goal of neuroscience: to become the most ambitious and altruistic initiative humanity has ever devised (Kandel, 2013). The following sections present some of the most challenging ethical concerns related to brain implants and their day by day more frequent use.

The challenges of brain implants

The main goal of brain implants is to make the life of people better, simpler and more dignified. Many patients who have a severe health problem might benefit from the deep brain stimulation and neural devices. While the implanted neuro-stimulators responsible for controlling and blocking abnormal nerve signals of patients with Parkinson’s disease are officially recognized and undergone by almost 140000 patients (Palferman, 2016), other implants have not been recognized and approved yet.

So, the first questions are:

- Who can be a good candidate for a BCI treatment (depending on the: level of the abnormal neural function; patient's age and condition; previous treatments; likelihood of improvement; and the potential counter indications);
- Who has the right to undergo the treatment (depending on the: social; religious; ethnographic or even a gender status);
- Who can hold the responsibility of the treatment (depending on the: level of consciousness of the patient: himself, the family, the doctors, the researchers; and the degree of unexpected consequences);
- Who holds the responsibility of unpredicted reactions of the patients, for example, a sudden extreme mood disorder, or a car accident caused by the patient with implanted neural devices, as suggested by Klein et al. (Klein, 2015);
- Could the brain implants affect the mental competence of the patient's personality resulting "in damage caused by undesirable or even deviant behavior", according to Klaming and Haselager (Klaming, 2013).

While the first two questions are common for any medical treatment, the responsibility issues are unquestionably technology dependent.

Assuming that the patient is conscious to request a BCI treatment, can afford it, and there are no evident signs indicating that the treatment is deteriorating other vital functions, is there any proof that the neural implants are safe, and that they can't trigger an instant paralyses, stroke, and even death of the patient? If such a circumstance occurs when the patient is performing a safety critical task, then many other people will also suffer.

Unfortunately, there are too many examples proving the unsafety issues of brain stimulators: various examples of unwanted risks due to average breaking up of the excessive synchrony, lowering of beta power and an average reduction in clinical impairment (Klein, 2013); poor anatomic accuracy and uncertainties (Calabrese, 2016); the risk of infection and hemorrhage (Glannon, 2014); followed by the observed and revealed hardware discomfort, and the necessity of surgical revision (Fenoy, 2014).

Once the safety of the brain devices is assured, and all unwanted side effects are bypassed, still their durability, reliability and robustness are the potential weak points. In his article, Walpaw mentions that current brain-computer interface research is challenged by various problems connected with the signal-acquisition hardware, and enlarged with the lack of convincing clinical validation, brain-computer interface dissemination and support (Walpaw, 2012). Therefore, the three critical areas suggested by Shih and his colleagues (Shih, 2012): signal-acquisition hardware; brain-computer interface validation and dissemination; and reliability should be seriously reconsidered.

It has been recently proved that brain-computer interfaces enhance the episodic memory (Burke, 2013). New technological trends go even further, they intend to enable synthetic telepathy (or, silent talk) and silent communication, turning "people into living multimedia machines" (Synthetic Telepathy).

Direct communication between human brain and computer external devices, as well as brain-to-brain interface via Internet are persistently monitored (Grau, 2014). Brain waves are wirelessly controlled, and consequently, neural implants are vulnerable to network attacks (Engber, 2016). Therefore, brain implants can be easily hacked and maliciously interfered (Pycroft, 2016).

The researchers involved in all the projects are aware that bio-electronic implants are "a potential threat to human dignity". These problems lead to a new phenomenon, called brain-jacking, the situation when attackers establish an unauthorized access to implants in order to manipulate with the patients (Pycroft, 2016). Under the "brain-jacking siege", security and privacy will no longer exist.

Conclusion

Since 1924, when Berger discovered the electrical activity of the human brain (Brazier, 1961), brain implants have been researched by many scientists. Berger's invention proved that one day, it will be possible to act through brain signals. Initially, brain-computer interfaces were focused on neuro-prosthetics applications intended for people with special needs. As mentioned earlier in this paper, recent studies have been extended to experiments directed to stimulate telepathy, and to enhance the memory. And, what is even more exciting, they start being used for nonmedical purposes, including education, brain controlled art, lie detection, game industry, and entertainment (Rao, 2013).

Brain implants will soon become pervasive, and people will start implanting them to become more intelligent, healthier, stronger, or to live longer. Whether the researchers who contributed to their development like it or not, humanity is slowly, but surely entering the cyborg era. Exactly fifteen years ago, Andy Clark suspected that we are natural-born cyborgs (Clarks, 2001). Two years ago, Nick Kolakowski tried to convince us that "We're already cyborgs" (Kolakowski, 2014). He was provoked to such an allegation as a response to Google glasses, Snowden's whistleblowing allegations, and Elon Musk's interview for Forbes. By then, science made many steps forward. The future, if we ever witness it, will show whether they will make the world a better place, or may be the pessimistic forecast that humans will no longer rule the world we know will unfortunately prevail. Similar dilemmas existed about the misuse of many technological inventions. So far, humanity was clever enough to overcome them.

References

Brazier, M. AB. 1961. A history of the electrical activity of the brain: The first half-century.

Burke, J. F., Merkow M. B., Jacobs J., Kahana M. J., and Zaghloul K. A. 2013. Brain computer interface to enhance episodic memory in human participants. *Frontiers in human neuroscience 8:* 1055-1055.

Calabrese, E. 2016. Diffusion Tractography, *Deep Brain Stimulation Surgery: A Review. Frontiers in neuroanatomy*, 10.

Clark, A. 2001. Natural-born cyborgs? *Cognitive Technology: Instruments of Mind*, pp. 17-24. Springer Berlin Heidelberg.

DARPA. 2013. DARPA and the Brain Initiative, *http://www.darpa.mil/program/our-research/darpa-and-the-rain-initiative*

EC. 2007 - 2012. Brain research, *http://ec.europa.eu/research/health/pdf /brain-research_en.pdf*

Engber, B. 2016. The Neurologist Who Hacked His Brain – And Almost Lost His Mind", *https://www.wired.com/2016/01/phil-kennedy-mind-control-computer/*

Fenoy, A. J., and Simpson Jr, R. K. 2014. Risks of common complications in deep brain stimulation surgery: management and avoidance: Clinical article. *Journal of neurosurgery*, 120(1), 132-139.

Glannon, W. 2014. Ethical issues with brain-computer interfaces. *Frontiers in systems neuroscience, 8*, 136.

Grau, C., Ginhoux, R., Riera, A., Nguyen, T. L., & Chauvat, H. 2014. Conscious Brain-to-Brain Communication in Humans Using Non-Invasive, *http://www.kurzweilai.net/first-brain-to-brain-telepathy-communication-via-the-internet*

Hassanien, A. E., and Az Ar, A. A. T. A. 2015. Brain-Computer Interfaces. *Springer International Publishing*

Joy, B. 2000. Why the Future Doesn't Need Us? *http://www.wired.com/2000/04/joy-2/*

Kandel, E. R., Markram, H., Matthews, P. M., Yuste, R., and Koch, C. 2013. Neuroscience thinks big (and collaboratively). *Nature Reviews Neuroscience*, 14(9), 659-664.

Klaming, L., and Haselager, P. 2013. Did my brain implant make me do it? Questions raised by DBS regarding psychological continuity, responsibility for action and mental competence. *Neuroethics, 6(3)*, 527-539.

Klein, E., Brown, T., Sample, M., Truitt, A. R., and Goering, S. 2015. Engineering the Brain: Ethical Issues and the Introduction of Neural Devices. *Hastings Center Report, 45(6)*, 26-35.

Kolakowski, N. 2014. We're already cyborgs, *http://insights.dice.com/2014/01/14/were-already-cyborgs/*

Kurzweil, R. 2001. The Law of Accelerating Returns, *http://www.kurzweilai.net/the-law-of-accelerating-returns*

Kurzweil, R. 2005. The singularity is near: When humans transcend biology. *Penguin*

Lebedev, Mikhail. 2014. Brain-machine interfaces: an overview. *Translational Neuroscience 5*, no. 1: 99-110.

Palferman, J. 2016. Deep Brain Stimulation: You Aint Seen Nothing Yet? *http://www.journalofparkinsonsdisease.com/blog/palfreman/deep-brain-stimulation-you-aint-seen-nothing-yet*

Pycroft, L., Boccard, S. G., Owen, S. L., Stein, J. F., Fitzgerald, J. J., Green, A. L., and Aziz, T. Z. 2016. Brainjacking: implant security issues in invasive neuromodulation. *World neurosurgery*.

Rao, Rajesh PN. 2013. Brain-computer interfacing: an introduction. *Cambridge University Press*

Rupp, R., Kleih, S. C., Leeb, R., Millan, J. D. R., Kübler, A., and Müller-Putz, G. R. 2014. Brain–computer interfaces and assistive technology. *Brain-Computer-Interfaces in their ethical, social and cultural contexts* Springer Netherlands. pp. 7-38.

Shih, J. J., Krusienski, D. J., and Wolpaw, J. R. 2012. Brain-computer interfaces in medicine. *Mayo Clinic Proceedings, Vol. 87, No. 3*, pp. 268-279. Elsevier.

Synthetic telepathy: Our mission, *http://www.synthetictelepathy.net/our-mission/*

Van Erp, J. B., Lotte, F., & Tangermann, M. 2012. Brain-computer interfaces: beyond medical applications. *Computer-IEEE Computer Society*, 45(4), 26-34.

Vinge, V. 1993. The coming technological singularity: How to survive in the post-human era. *http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022856.pdf*

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. 2002. Brain–computer interfaces for communication and control. *Clinical neurophysiology, 113(6)*, 767-791

Wolpaw, J. R. 2012. Brain-computer interfaces: progress, problems, and possibilities. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pp. 3-4. ACM.

To teach a machine a sense of art – problems with automated methods of fighting copyright infringements on the example of YouTube Content ID

Justyna Zygmunt
Chair of Intellectual Property, Jagiellonian University in Kraków, Poland
justyna.zygmunt@uj.edu.pl

Abstract

The goal of this article is to show the delicate trade-off between safeguarding the fair use provisions of the copyright law and effective forms of reacting upon copyright infringement. In the light of the number of Internet users, the major hosting providers are basically unable to manually verify infringement notices and need to employ automated forms of filtering content. Unfortunately, these tools don't detect fair use, at least not at the moment. The questions are: can we teach the machines how to detect the fair use? How to do it? And in case of negative answer, should we put the priority on the enforcement or on the fair use provisions?

Factual and legal background

In the era of Web 2.0, when the vast majority of Internet users are also the creators of content (user generated content, UGC), the number of copyrightable works has remarkably increased. Every minute Internet users create about 200 000 new Instagram photos, 250 000 tweets, and 70 hours of YouTube videos, most of which is eligible for protection. A great share of the content is being stored at major hosting providers like YouTube (hosts, service providers, providers, SP). In the context of the flood of creation, hosting providers are unable to hire enough manpower to verify whether the UGC remains in line with copyright law. The inability to check whether the UGC infringes copyright law opened a discussion on secondary liability of providers that made the Internet business look very risky from providers' perspective. In order to make the online storage of content legal and possible, law releases SP from liability for copyright infringements that occur within their services, if they comply with certain requirements (safe harbours). In the U.S. this rule is introduced by the Digital Millenium Copyright Act, in Europe it's the E-Commerce Directive 2000/31/EC.

These regulations introduce a rule according to which service provider is not liable for the information stored at the request of a recipient of the service, on condition that the provider does not have actual knowledge of illegal activity or information and, as regards claims for damages, is not aware of facts or circumstances from which the illegal activity or information is apparent or the provider, upon obtaining such knowledge or awareness, acts expeditiously to remove or to disable access to the information (notice & takedown). SP are not obliged to monitor their own websites for infringing content, but only have to immediately react upon a notice.

YouTube

YouTube is a major hosting provider that offers a possibility to store videos. According to the statistics[1], it has over a billion users – almost one-third of all people on the Internet – and every day people watch hundreds of millions of hours on YouTube and generate billions of views. Due to tremendous number of users and potential copyright infringement claims, YouTube – as one of the first websites – launched an automated filtering system called Content ID, which supports the notice & takedown procedure.

Statistics indicate that hundreds of millions of videos have been claimed using this tool since its inception. Content ID made the process of reacting upon infringements faster and cheaper than traditional notices filed in courts. Many commentators argue that such systems are inevitable future of copyright enforcement.

[1] https://www.youtube.com/yt/press/statistics.html

Content ID

YouTube's Content ID system is a landmark example of automated filtering system. Copyright holders who have Contend ID upload copies of their works to YouTube that digitizes works and creates fingerprints of them. Those fingerprints can be compared with other videos uploaded to the service and when the match is found, the system flags the content. Once a video has been flagged as containing copyrighted content, the system sends a notification to the copyright holder who can take an action against the presumed infringer. This system favours mass corporations that can easily deter single users from creating new content partially built upon already existing works.

Problem

According to the copyright law, not every form of using someone else's work constitutes an infringement. Some of them are legitimized on the grounds of the doctrine of fair use. In order to fall within the scope of this doctrine, a particular use must fulfil certain conditions (e.g. must be carried out for educational purposes). The conditions are usually described in a vague and general manner in order to allow a flexible, case-by-case analysis and are subject to varying methods of interpretation. The problem with automated filtering systems like Content ID is that the machines don't detect fair use.

Detecting fair use is a highly complicated task even for experienced copyright lawyers. It requires various assessment: if the original work is a "work" within the meaning of copyright law, if there is an inspiration or a derivative work and finally, if the requirements of the fair use are met. The latter part is the hardest, especially in the Anglo-American systems where knowledge on art and economy is required to make a full analysis. What is more, the conditions for fair use differ from country to country and are – because of the changing judiciary – in a constant flux.

At the moment, according to my best knowledge, hosting providers don't even attempt to make a system that will meet fair use provisions. They base the infringement notices simply on the fact that one work appears within another and don't consider additional factors. Some of them introduce strict rules e.g. taking up to 5 seconds of someone else's work is allowed, but such rules have nothing to do with the traditional understanding and purpose of vague fair use provisions. The most famous example of a borderline fair use-YouTube case is Lenz vs. Universal Music Corp (so called "dancing baby case"); it will be briefly described during my presentation.

The current state of things, where major hosting providers use machines to filter the content and base the infringement notices on the simple fact using an excerpt of someone else's work, impairs the freedom of speech of Internet users who can be easily banned from criticizing someone else's works, and contradicts one of the very basic goals of copyright law which is encouraging creation.

Questions

Should we agree with the current state of things where the fair use provisions are pushed into the background in order to provide an effective enforcement? Can we replace tons of books and tens of thousands of copyright attorneys by "simply" teaching machines about boundaries of using someone else's works and if so, how to do it? Should the fair use guidelines be simplified or the algorithms more advanced? Who should be authorised to prepare these instructions? How to make them up-to-date? Generally: how to strike a balance between fighting Internet pirates and providing an environment encouraging freedom of speech and expression and how AI can help to achieve this goal.

My presentation aims to provide answers to these questions.

References

D. E. Ashley, *The Public as Creator and Infringer: Copyright Law Applied to the Creators of User-Generated Video Content*, 20 Fordham Intell. Prop. Media & Ent. L. J. 563, 572-73 (2010);

R. Barber, *Viacom v YouTube: Is YouTube paddling out of the safe harbour?,* Ent. L.R. 2012, 23(7), 220-224;

Z. Carpou, *Robots, pirates, and the rise of the automated takedown regime: using the DMCA to fight piracy and protect end-users*, 39 Colum. J.L. & Arts 551, Summer 2016;

C. Hui Yun Tan, *Lawrence Lessig v Liberation Music Pty Ltd - YouTube's hand (or bots) in the over-zealous enforcement of copyright*, E.I.P.R. 2014, 36(6), 347-351;

C. Hui Yun Tan, *Technological "nudges" and copyright on social media sites,* I.P.Q. 2015, 1, 62-78;

D. Halbert, *Mass Culture and the Culture of the Masses: A Manifesto for User-Generated Rights*, 11 Vand. J. Ent. & Tech. L. 921, 936 (2009);

L. Leister, *YouTube and the Law: A Suppression of Creative Freedom in the 21st Century*, 37 T. Marshall L. Rev. 109, 118 (2011);

D. Manna, *Artificial Intelligence Insourcing: Why Software Technology Will Dominate Legal Process Outsourcing for Routine Document Drafting*, Canadian Journal of Law and Technology, 12 Can. J. L. & Tech. 109, June 2014;

P. McKay, *Culture of the Future: Adapting Copyright Law to Accommodate Fan-Made Derivative Works in the Twenty-First Century*, 24 Regent U. L. Rev. 117, 124-27 (2011);

L. Solomon, *Fair users or content abusers? The automatic flagging of non-infringing videos by Content ID on Youtube*, Hofstra Law Review, 44 Hofstra L. Rev. 237, Fall 2015.