(Technical) Autonomy as Concept in Robot Ethics

Michael Funk(✉) and Mark Coeckelbergh

Department for Philosophy of Media and Technology, University of Vienna, Vienna, Austria
funkmichael@posteo.de, mark.coeckelbergh@univie.ac.at
http://www.funkmichael.com,
http://www.coeckelbergh.wordpress.com

Abstract. This paper aims to contribute to the debate about ethical, legal, and social implications of robotics by discussing the meaning of autonomy. Robots are often labeled as autonomous, but what does "autonomy" in robotics actually mean? In order to answer this question, methods of conceptual analysis and language critique are applied. It turns out that the empirical-descriptive application of the word autonomy in technical context is different to the normative usage of autonomy in human life and human societies. Following this insight, and an embodied approach in philosophy of technology, six forms of technical tools are briefly introduced which could be used to describe several levels of technical autonomy. The different forms are summarized in a heuristic scheme, which can be used to set up clearer applications of the word "autonomy" in ethical, legal and social debates concerning robotic technologies.

1 Introduction

Robot ethics is a field of technology ethics with a genuine focus on robotic systems. But robots as moral or ethical agents also plays a role in the debate. Since there is no perfect definition of robot, no final decision of what robot ethics means has been established yet. But some working definitions are available and serve as tentative practical starting point for debates in robot ethics. E.g.:

The etymological *locus classicus* by Josef and Karel Čapek, who defined "robot" in their play "Rossum's Universal Robots (R.U.R.)" in 1920/1923 as a forced labourer (from Czech "robota" = forced or compulsory labour, corvée) [1].

Or a current EU working definition in the *EUROPEAN CIVIL LAW RULES IN ROBOTICS, Oct. 2016*: "Paragraph 1 of the motion for a resolution suggests that the Commission 'propose a common European definition of smart autonomous robots and their subcategories by taking into consideration the following characteristics of an intelligent robot: acquires autonomy through sensors and/or by exchanging data with its environment (inter-connectivity) and trades and analyses data; is self-learning

This paper has been financially supported by the EU Horizon 2020 INBOTS project (grant agreement No. 780073), working package 2 (debate on legal, ethical and socio-economic aspects).

J. L. Pons (Ed.): INBOTS 2018, BIOSYSROB 25, pp. 59–65, 2020.
https://doi.org/10.1007/978-3-030-24074-5_12

(optional criterion); has a physical support; adapts its behaviours and actions to its environment'" [2].

Two influential authors who shaped the notion of robot ethics in current debates are Gianmarco Veruggio and Keith Abney. They define robot ethics as a concept with three layers of meaning:

Robot ethics as applied ethics;

Robot ethics referring to the moral code/morality that might be programmed into machines;

Robot ethics meaning that robots themselves perform ethical reasoning [3].

This paper aims to contribute to the discussion of robot ethics as applied ethics (meaning 1). Since applied ethics is more about pragmatically successful moral practice than abstract ethical theories, a special focus is on terms which represent the meaning of new technological capabilities. Clarifying the conceptualization of foundational terms can contribute a rational pragmatic discourse about the application of robotic systems. Therefore, the term "autonomy," which is used in the second definition, will be emphasized in this short paper.

Another key term is risk. Whenever we apply new technologies we are confronted with risky situations, a certain probability of accidents, and a lack of empirical knowledge [4, 5]. What do we know and what could we know about the consequences of our actions? How can we assess the environmental, political or social risks of robots? In robot ethics the risk debate is closely related to the concept of responsibility [6]. Who is responsible for a risky robotic system that fails to fulfill its intended function and causes damage: The robot, the user, the engineer or the person who sells the system? Risk and responsibility relate to technical capabilities and features of robotic systems which are discussed as autonomy. So called "autonomous" robots are able to interact with the environment independent of any remote control. This issue directly relates to the question of risk and responsibility. On the other hand, using terms like responsibility or autonomy in order to describe robots is problematic. Autonomy and responsibility are terms developed for human agents in human societies. Using human terminology for the description of technical functions leads to a linguistic anthropomorphism. We describe robots as if they would be autonomous or might even be responsible, and in consequence we address implicitly human values to machines. By doing so we might end up committing methodological mistakes like mixing up terms that describe a matter of fact and terms that describe what ought to be (see Hume's law and the naturalistic fallacy described by G.E. Moore [7, 8]). It is important to make sure that there is a basic difference between the normative meaning of "autonomy" in human life and the empirical descriptive meaning of "autonomy" in technical praxis.

With this short paper we don't want to present final solutions for the epistemic, conceptual and ethical problems of autonomy, responsibility and risk in robotics. Our aim is to contribute a concrete suggestion to a better understanding of the empirical descriptive meaning of (technical) autonomy; and therefore also to the investigations of the INBOTS consortium in WP 2 by presenting conceptual analysis of one basic term in the debate. Following a language critical, hermeneutical and constructivist approach in the tradition of Ludwig Wittgenstein [9] and others, we are going to sketch out a template and visual heuristic scheme which shows similarities and differences between

the several layers of meaning. By systematically clarifying the different layers of meaning of the terms autonomy, responsibility and risk, we aim to present some (at least tentative) conceptual analysis that can contribute to further detailed investigations in order to "promote debate[s] on legal, ethics & socio-economic aspects" also with an interdisciplinary intention.

2 Six Forms of (Technical) Autonomy

Following the approach of methodological constructivism and culturalism (Methodischer Konstruktivismus und Kulturalismus) [10], the language of everyday life can be seen as a methodological starting point for more complex scientific conceptualizations. Peter Janich argues that, because of its multi-perspective situatedness within intentional actions, human communicative competences are principally not substitutable by technical systems. There is a fundamental difference between the praxis of communication with all its gestural and tacit ways of socially shared interactions and technical transmission of pure disembodied information [11–13]. Applying this argument to information technologies we claim not to understand, for instance, (social) robots as "autonomous," "intelligent" or "creative" actors, but as technical tools and aspects of means-end oriented social practice. Within these social practices we humans play with words like "autonomy" on the basis of our everyday life experience. But this does not mean that robots in a normative sense are free in their actions and therefore can be attributed as "autonomous" (in the classical Kantian meaning). However, those tools can be considered as parts of techno-material cultures – even if robots are technically more complex means than hand axes or hammers [14]. Six forms of technical tools can be summarized: 1. handcraft, 2. machine, 3. automat, 4. embedded technical autonomy, 5. technical semi-autonomy, and 6. autonomy.

A differentiation between these six forms is enabled by the categories "energy," "movement/process," and "control/framework." Form 1 to 3 belong to "pre-modern & modern technologies" including early handcraft instruments, wooden and stone tools like hand axes (form 1), which are fully controlled and applied by human actors with sensorimotor skills. Hereby the energy, movement – including routine, as well as new ways of usage –, and the framework – including the end and monitoring – of a technical action are fully provided by the human body. Success of a technical performance while using hand axes depends on the sensory capacities of a person while using the tool. Form 2 and 3 include e.g. weaving machines or excavators – tools where the energy afforded for technical success comes from the artefact, and aspects of routine are implemented as well. Praxis such as problem solving and finding creative solutions for various contingent and unintended situations, setting the aims of a technical procedure, as well as the monitoring of action all remain embedded in the human body.

Form 4 and 5 are related to "hypermodern technologies", including computers and robotics. The term hypermodern (e.g. Albert Borgman [15]) means that modern, industrial technologies are not redeemed by postmodern, 20th-century developments: moreover, when it comes to technologies, modern developments are moreover boosted and enhanced in the 20th and 21st century. "Hypermodern" in this sense means not a historical cut but a further step on the developmental path that started in early

modernity, 16th- and 17th-century technosciences. Social robots belong to form 4 and 5, not to form 6, which is a postulate or a posit of something that eventually becomes real: a tool that is totally autonomous. In such a case it might be adequate to avoid the word tool (form 6). But current robots and technical systems will not belong to this hypothetical category in the foreseeable non-science fiction future.

What is the difference between form 4 and 5 on the one hand and form 6 on the other? In contrast to form 1 through 3, in the forms 4 and 5 aspects of praxis, problem solving and the setting of aims are carried out by the technical system. The success of a technical practice, especially in form 5, is more independent from the human body. Human actors remain in the position of surveillance and intervention in case of a functional defect. For this reason, form 4 is called "embedded technical autonomy": to make sure that some aspects of creative problem solving in contingent situations are emulated (not a 100% copy of human creativity or autonomy!) in the technical tool. As the means-end setting e.g. of social robots is generated by human actors, this category of technical tools includes the word "embedded". It is a technical form of autonomy which enables capacities of social interaction – e.g. giving a spontaneous verbal reaction to an unintended sentence of a user. But for epistemic reasons this is not the same as human autonomy.

Moreover, in form 5 tools like social robots functionally capture parts of the framework (the "embedding" of form 4). Here the system sets some aims of its own functionality – always under monitoring of human actors: e.g. when social robots include complex user profiles based on manifold sensor data. The robot starts "learning" about its environment and thereby enables capacities for functionally finding own aims. For instance, when a user of a social robot often forgets his house door key, the robot might include this issue in the user profile and independently start some games or exercises with the user in order to train his skills in memory and attentiveness. Again, surveillance and intervention in case of dysfunction are related to human bodily actors. This form 5 is called "technical semi-autonomy" as it is not a replacement of human intelligence, creativity or autonomy, but a technical functionality which includes the emulation of some means-end capacities. Form 6 in contrast is the hypothetical postulate of total autonomy in a technical entity. Here all (possible) sensory layers and cognitive domains of human bodies would be represented in the "tool." This is science fiction, but not totally inconceivable.

Terminologies and concepts of technical autonomy or are treated in ongoing and controversial discussions (e.g. [1, 16, 17]). The six forms presented here are one possible interpretation of a plausible approach to differentiated forms of technical tools and their related notions of (technical) "autonomy." The following table illustrates these six forms and should summarize the above mentioned differences between the tool hand axe and the tool social robot (as shown in Fig. 1, source: [18]).

It would be interesting and intellectually profitable to further develop this approach including the several meanings of "autonomy," "risk" and "responsibility" – also within the WP 2 of the INBOTS project. But in the scope of this short paper it is more important to see that (social) robots are technical tools (form 4 and 5) and to further focus on the epistemic ways in which these technical systems/tools have an impact on alterity relations and mediate human communication. Social robots are therefore interpreted as aspects of techno-material culture, where robots mediate sensory

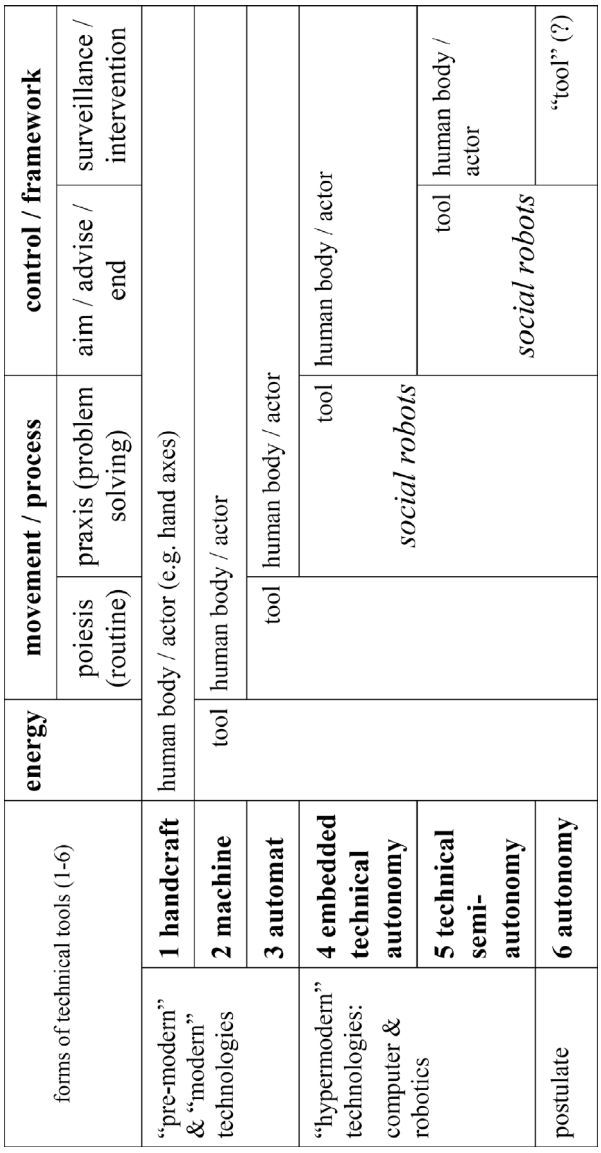

Fig. 1. A schematic summary of six forms of technical tools and technical autonomy.

perception and bodily practice in particular *situations*. Whole situations generate frameworks for the interpretation of practice on the context of human finitude and bodily as well as social vulnerability [19]. There is no action isolated from its context. But the contexts also differ; no master-situation can be defined. This methodological impact has been philosophically investigated as apriori-situation (*Situationsapriori*) [20, 21]. Knowledge research includes non-reducible forms of knowing that epistemologically can be brought into heuristical schematic form in order to create something

like a philosophical toolbox [22]. With a *heuristics* the epistemological and multi-perspective range of these mediations can be brought in the form of a lucid depiction (*übersichtliche Darstellung*). Ludwig Wittgenstein developed in his late writings a methodology of philosophical depiction. Philosophy becomes a practice itself and investigates the – tacit, meaningful – grammar of bodily actions [23]. Lucid depiction (*übersichtliche Darstellung*) becomes hereby one possible tool for clarifying the many ways in which we use words in a meaningful way [9, 24].

3 Conclusion and an Ethical Outlook

In this short paper we briefly summarized six forms of technical tools. Our starting point was a conceptual analysis and a bodily understanding of technical praxis. The more aspects/features of human bodily actions are included into a tool, the more "autonomous" it becomes. It is important to make sure that this refers to an empirical descriptive usage of the word autonomy which might entail normative implications, although it should not be mixed up with the genuine normative usage of autonomy in human social life. This paper contained an analysis of the first meaning: empirical descriptive application of "autonomy" in order to describe the application of technical tools. Additionally, conceptual analysis of "risk" and "responsibility" can be added in future steps. With respect to the INBOTS project, a genuine emphasis on the genuine normative meaning of "autonomy" in social life is necessary as well. INBOTS means "Inclusive Robotics for a better Society." Since this slogan represents the empirical descriptive side of the coin, it must also be turned around in order to critically and ethically analyze the human autonomy in the loop: "Better Robotics for an inclusive Society." What is a better society? Which life do we want to life? How can "autonomous" robots support human autonomy and inclusion, instead of replacing it? These are some of the questions which should be further discussed also in the context of so called "autonomous robots," related risks and responsibility.

Acknowledgment. We like to kindly express our thanks to our colleagues involved in the INBOTS project for the critical discussions, feedback and creative brainstorming processes at the Pisa conference in October 2018, but also in previous meetings, especially Daniel López Castro, Fiachra O'Brolchain, Maria Amparo Grau Ruiz and Mario Toboso.

References

1. Edgar, S.L.: Morality and Machines. Perspectives on Computer Ethics. Second Edition, p. 455. Jones and Bartlett, Sudbury (2003)
2. http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf. Accessed 10 Sept 2018
3. Veruggio, G., Abney, K.: Roboethics: the applied ethics for a new science. In: Lin, P., Abney, K., Bekey, G.A. (eds.) Robot Ethics. The Ethical and Social Implications of Robotics, pp. 347–363. MIT Press, Cambridge/London (2012)
4. Nida-Rümelin, J.: Ethik des risikos. In: Nida-Rümelin, J. (ed.) Angewandte Ethik. Die Bereichsethiken und ihre theoretische Fundierung. Ein Handbuch. 2., aktualisierte Auflage, pp. 862–885. Kröner Verlag, Stuttgart (2005)

5. Ropohl, G.: Verantwortung und risiko. In: Heidbrink, L., et al. (eds.) Handbuch Verantwortung, pp. 887–908. Springer, Wiesbaden (2017)
6. Lenk, H., Maring, M.: Verantwortung in technik und wissenschaft. In: Heidbrink, L., et al. (eds.) Handbuch Verantwortung, pp. 715–731. Springer, Wiesbaden (2017)
7. Hume, D.: Ein Traktat über die menschliche Natur. In: der Grundlage der, A., von Theodor Lipps, Ü., herausgegeben von, n., Brandt, H.D. (eds.) Teilband 2. Buch II Über die Affekte. Buch III Über Moral, pp. 546–547. Felix Meiner, Hamburg (2013)
8. Moore, G.E.: Principia Ethica. Erweiterte Ausgabe, pp. 40–44. Reclam, Stuttgart (1996)
9. Coeckelbergh, M., Funk, M.: Wittgenstein as a philosopher of technology: tool use, forms of life, technique, and a transcendental argument. Hum. Stud. **41**(2), 165–191 (2018)
10. Funk, M., Fritzsche, A.: Engineering practice from the perspective of methodical constructivism and culturalism. In: Michelfelder, D.P., Doorn, N. (eds.) Handbook of Philosophy of Engineering. Routledge, Abingdon (forthcoming)
11. Janich, P.: Substitution kommunikativer Kompetenz? In: Decker, M. (ed.) Robotik. Einführung in eine interdisziplinäre Diskussion, pp. 17–31. Graue Reihe, Bad Neuenahr-Ahrweiler (1999)
12. Janich, P.: Kultur und Methode. Philosophie in einer wissenschaftlich geprägten Welt. Suhrkamp, Frankfurt a.M. (2006)
13. Janich, P.: Was ist Information? Kritik einer Legende. Suhrkamp, Frankfurt a.M. (2006)
14. Funk, M.: Humanoid robots and human knowing – perspectivity and hermeneutics in terms of material culture. In: Funk, M., Irrgang, B. (eds.) Robotics in Germany and Japan. Philosophical and Technical Perspectives, pp. 69–87. Peter Lang, Frankfurt am Main a.o. (2014)
15. Borgmann, A.: Technology and the Character of Contemporary Life: A Philosophical Inquiry. The University of Chicago Press, Chicago/London (1984)
16. Lin, P., Abney, K., Bekey, G.A. (eds.): Robot Ethics: The Ethical and Social Implications of Robotics. MIT Press, Cambridge/London (2012)
17. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, Oxford (2009)
18. Funk, M.: Paleoanthropology and social robotics. Old and new ways in mediating alterity relations. In: Aagaard, J., Friis, J.K.B., Sorenson, J., Tafdrup, O., Hasse, C. (eds.) Postphenomenological Methodologies. New Ways in Mediating Techno-Human Relationships, pp. 125–149. Rowman & Littlefield/Lexington (2018)
19. Coeckelbergh, M.: Human Being @ Risk: Enhancement, Technology, and the Evaluation of Vulnerability Transformations. Springer, Dordrecht (2013)
20. Rentsch, T.: Die Konstitution der Moralität. Transzendentale Anthropologie und praktische Philosophie, p. 68ff. Suhrkamp, Frankfurt a.M. (1999)
21. Rentsch, T.: Heidegger und Wittgenstein. Existenzial- und Sprachanalysen zu den Grundlagen philosophischer Anthropologie, pp. 75ff. Klett-Cotta, Stuttgart (2003)
22. Abel, G.: Knowledge research: extending and revising epistemology. In: Abel, G., Conant, J. (eds.) Rethinking Epistemology, vol. 1, pp. 1–52. De Gruyter, Berlin/Boston (2012)
23. Wittgenstein, L.: Philosophische Untersuchungen. In: Wittgenstein, L. (ed.) Werkausgabe Band 1. Tractatus logico-philosophicus. Tagebücher 1914–1916. Philosophische Untersuchungen, pp. 225–577. Suhrkamp, Frankfurt a.M. (2006)
24. Gabriel, G.: Logisches und analogisches Denken. Zum Verhältnis von wissenschaftlicher und ästhetischer Weltauffassung. In: Demmerling, C., Gabriel, G., Rentsch, T. (eds.) Vernunft und Lebenspraxis. Philosophische Studien zu den Bedingungen einer rationalen Kultur, pp. 157–174. Suhrkamp, Frankfurt a.M. (1995)