Cross-cultural Application of Ethical Principles in the Design Process of Autonomous Machines

Anniina Huttunen
Institute of International Economic Law (KATTI), University of Helsinki
Helsinki, Finland

William Brace
Department of Engineering Design and Production, Aalto University - School of Science and Technology
Espoo, Finland

Vesa Kantola
Department of Media Technology, Aalto University - School of Science and Technology
Espoo, Finland

Lorenz Lechner
Central Facility of Electron Microscopy, University of Ulm
Ulm, Germany

Jakke Kulovesi
Department of Automation and Systems Technology, Aalto University - School of Science and Technology
Espoo, Finland

Kari Silvennoinen
Department of Business Technology, Aalto University, School of Economics
Helsinki, Finland

ABSTRACT

Engineering is defined as an activity that aims at producing useful things that generate human benefit. However, currently a large part of the robotics industry is driven by the development of "killer applications" capable of causing tremendous amount of human suffering and harm. We propose a twofold solution to the ethical dilemma: external ethical guidelines combined with intrinsic engineering practices. As a first step to help mitigate the anticipated problems, governments and international organizations should promote a generally accepted codification of roboethics. This codification should comprise a basic set of hard boundaries that must not be crossed under any circumstances. The first step can be seen as an external ethics approach. As a second step, we propose that the residual ethical risk should be taken into consideration by implementing an oath for technology developers (New Archimedes' Oath), analogous to the Hippocratic oath. We understand this oath as an internal ethical risk management tool that increases developer's awareness of ethical machine development while leaving an appropriate level of latitude for making individual decisions. Furthermore, we show how to implement the oath as a machine design principle. This hybrid approach can also be seen as a tool for cross-cultural application of ethical principles in the design process. We use robotics as an example case, but the oath concept goes beyond guidelines for specific development areas.

Keywords: Meta-Engineering Praxis, ethical design and ethical engineering, Robotics, ethical issues and social responsibility, organizational/societal impact, HRI design, Artificial Intelligence

1. INTRODUCTION

The debate about the ethical implications of robots is at least as old as the concept of robots itself. It has been the topic of countless fiction and non-fiction works. Lately there has been a new rise in the topic as robots are starting to become commonplace in more and more applications. Ethical risks have been considered especially in the area of care of children and the elderly, and in context of autonomous robot weapons [1].

Unfortunately, the problems related to the increased capability for autonomy are not solved by using simple rules like Asimov's laws[1]. First and foremost, it should be kept in mind

[1] Decades ago, science fiction writer Isaac Asimov developed three laws of robotics. According to those rules 1) a robot may not injure a human being or, through inaction, allow a human being to come to harm, 2) a robot must obey orders given to it by human beings, except where such orders would conflict with the First Law, and finally 3) a robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

that these rules do not have any legal validity. Asimov's laws are a product of science fiction, only later used as a starting point for ethics analysis. In addition to the lack of formal validity, there are some fundamental shortcomings in these fictional laws. In the first place, these "laws" are robot-centric and disregard the role of the designers of the intelligent machines [2]. Recently, the focus has been shifting towards system engineering approaches to tackle some of the ethical problems around autonomous machines.

We start this article by describing why robots cannot take responsibility for their actions (2. Autonomous Machines and the Limits of Robot Laws). Then, we continue by presenting a hybrid framework for implementing ethical behavior into autonomous machines (3. Beyond General Codes of Conduct: Improving Engineering Ethics by Archimedes' Oath). The approach complements external ethics as codified in international codes of conduct with a flexible set of internal ethics derived from design principles and engineering ethics. We discuss the promotion of this internal ethics by codifying it into Archimedes' Oath. Finally, we show how to implement such an oath within a Systems Engineering (SE) approach (4. System Design and Oath).

2. Autonomous machines and the limits of robot laws

Currently, a large part of the robotics industry is driven by the development of "killer applications". Unlike the common "killer application" in other industries, the word killer application in robotics can be taken quite literally. The military industry is one of the driving forces in the development of robot technology, shown by the active research on military robots' ethics [3], [4]. Even though these autonomous devices currently almost all still have a human operator in the decision making process, it is foreseeable that future devices will operate fully autonomously. When we cross this boundary it is no longer clear where the burden of ethical decision-making lies. Unsurprisingly, many of the problems we face in today's technological environment have been recognized previously in science fiction literature. One of these problems is the ethical dimension of robot development. Having the robot act as a moral agent commonly solved the problem. In reality however, robots cannot take responsibility for their actions. For various legal and practical reasons the ethical responsibility rests on somebody else, namely either on the manufacturer, owner, or developer of the machine.

Whereas Asimov's laws assign morality to the robot, an alternative is to approach the moral issues from the developer side, seeking solutions by encouraging good design practices. Arguably, the robot-centric approach faces seemingly insurmountable problems in subjectivity of ethics, implementability, acceptability, and robot moral accountability [5], [6]. No satisfactory formal system of universal ethics has been constructed. Universal and formal ethics constructions have eluded ethics researchers due to the immediate problems relating to both subjectivity of moral conceptions and to the practical impossibility of formalizing such constructs. Moral psychology research can give a broad overview to the topic [8] but no definite and quantifiable solutions. In essence, morality does not bend to a formal set of rules without becoming impossibly complex or losing its connection to the human judgment of morality. Even if a robot would be equipped with a practically satisfactory level of ethics processing capability, the question of acceptability remains. Humans would be unlikely to accept a robot as an independent ethical being but direct blame towards the manufacturer and designer levels. Similarly, accepting robots as morally accountable beings is not only difficult but also would require significant advances in AI to make any sense [7]. Accountability presupposes that the actor's cognitive capabilities are sufficient for it to realize the accountability and to consequently have internal tendency to behave responsibly. This vision is still very far in the future for robots regardless of their rapidly enhancing skills and capabilities.

Superficially any discussion treating robots as autonomous agents in their own right and on a similar level with human beings could be considered purely academic. While this is still mostly true today, courts have already recognized robot judgment superior to human judgment in certain circumstances [8], [9], [10]. Still, it has to be pointed out that the capability to develop artificial intelligence is not connected to civil rights. In other words, rationality is not the decisive point in circumstances when it comes to human rights. Even if the machines would work perfectly correct and perfectly intelligent in compliance with the Bayesian inference rules, the rights and responsibilities are considered to ultimately rest on humans. Humans can be obliged to comply with orders given by an aircraft autopilot, like in those court cases mentioned above, but at the same time robots are still considered property of humans. This makes it necessary to find a solution to the ethics dilemma that satisfies the requirements of all parties: manufacturers, developers, owners, and users. Furthermore, for the reasons discussed earlier a solution implemented outside the autonomous system is highly desirable.

3. Beyond general codes of conduct: improving engineering ethics by Archimedes' oath

In the previous chapter, we have discussed external constraints to engineering. In the case of autonomous machines for military applications, these constraints are given primarily by international rules for armed combat and by general ethics considerations.

Furthermore, there are also ethical norms specific to the engineering processes that influence the behavior of autonomous machines. This 'engineering ethic' is a specific kind of ethics in a sense that the value of the work is often instrumental. There is no fundamental intrinsic value, such as justice or health, but the results of this work aim at instrumental goals. These goals are often value dependent. [11] The goal of an engineer is to develop a machine, which works as it should work. A "good" machine in these terms is a machine that has the functions listed in the specifications. This approach works more or less well as long as the machine created is simply a tool that augments human capabilities. Then the moral responsibility for misuse of the tool rests ultimately with the human operator. Autonomous machines are different in this respect. They are better treated as agents that work together with humans instead of mere tools utilized by humans. This change shifts considerably more moral responsibility to the machine's developer on both the design and implementation level.

Addressing this shift is becoming more and more important since the complexity and potency of machines currently increases dramatically while their costs decrease. This means that machines are becoming an integral part of an increasing number of, if not even all, aspects of our daily lives. This in turn increases the potential harm machines can possibly inflict to individuals and society. For this simple reason, we should expect engineering ethics to grow in importance. The ever-accelerating process of creating machines on the limit of the technological possibilities should be balanced with a counter-force. Interestingly, engineering is not singular in this respect. Other domains of rapid progress are facing equal problems as well. Life sciences and genetic engineering in particular are research fields with inherent high risk, real and perceived. In those fields, ethics considerations have long been an essential part of development, also as a necessity to secure continuing support for public funding. The key aspect in this strive for ethic conformity has been transparency, enabling the general public to observe the methods and ways of doing science. In order to increase trust on scientific progress and encourage the public to use new tools and products, the information needs of the public have to be satisfied [12]. The scientific method and peer-reviewed journals are now widely accepted as the gold standard for scientific progress. We argue that a similar level of trust and transparency is required in the development of intelligent machines.

Engineering generally does not take place in an academic environment. Transparency is not always wanted in e.g. commercial and military research. However, producers are interested in having high public acceptance and being able to bring their robotic products to the market. This is where the leverage for implementing ethics comes from. Societies and their governments can regulate their markets. Furthermore, governments are also a large customer for autonomous machines. It can be argued that currently the autonomous robots market is driven by military applications, and thus by government or tax-payer spending. Ethics considerations could therefore be implemented as a system design requirement by the society. However, on which level should this ethics be installed? We already discussed that currently they cannot be implemented solely within the machines themselves. What about anchoring them on the project level? This seems unreasonably difficult because then every new project would have to undergo an ethics evaluation process checking its adherence to the appropriate standard down to the level of every detail. While this is in principle a good approach, it is not the most elegant solution. Also, while the general ethical nature of a project needs to be discussed broadly, often only experts can understand and decide about the minutia. This suggests that the ethics control should be with the people that are actually working on the development of autonomous machines: with the engineers. Agreeably, this still leaves the question on how to implement this in practice unanswered. Fortunately, there are examples where ethics has been implemented similarly in a process: medical doctors are performing a critical task and are confronted with ethical questions on a day-to-day basis. There, for more than two thousand years the adherence to ethics standards has been implemented by a personal codified commitment of every practitioner: the Hippocratic oath.

We think that for engineers, a similar type of oath could be the much-needed solution. It could fulfill all requirements that we discussed earlier as being essential for a risk mitigation tool. In particular, it is implemented right at the developer level and does not require unrealistic advances in AI technologies. Naturally, this idea is not new: in 2001 ICSU research “Standards for Ethics and Responsibility in Science – an Empirical Study”, a follow-up to the 1999 World Conference on Science and of the decisions of the UNESCO General Conference [13] suggested the following Archimedes oath:

The New Archimedes’ Oath - Institut National Polytechnique de Grenoble (2000)

1. I will practise my profession abiding by the ethics of human rights and I will be aware of my responsibility for mankind's natural heritage.

2. In all acts of my professional life I will assume my responsibility towards my institution, towards society and towards future generations.

3. I will pay special attention to promoting fair relations between all men and supporting the development of economically underprivileged countries.

4. I commit myself to explaining my choices to decision-makers and citizens, making these choices as transparent as possible.

5. I will give priority to the forms of management permitting broad co-operation between all the actors with a view to making everyone's work and innovations meaningful.

6. I pledge myself to respecting ethical codes as well as examining and using means of information and communication critically.

7. I will take special care to honing my professional skills in all aspects of technological, economic, human and social sciences involved in my work.

What makes this oath a risk mitigation tool? When discussing risk related to autonomous machines it is necessary to distinguish two different causes: risk stemming from intentional "malicious" behavior and risk from faulty behavior. The New Archimedes' Oath addresses both risk categories. The oath's items 1 through 3 are clearly focused on reducing the intrinsic risk by connecting the engineer to external ethics frameworks. Item 7 is addressing how to improve the quality of the work in general. This point is critical in reducing residual risk from machine malfunction. Items 4 through 6 are not directly related to risk mitigation but are means of promoting the acceptability of the engineering endeavor and the oath itself. In the context of autonomous machines these points create confidence in the design process, which is likely to also extend to the resulting product. Seen as a whole, the oath seems to be an excellent lever for introducing the required internal ethics component. In particular, it has the great advantage to be a tried and tested highly successful tool in other disciplines.

Unfortunately, in engineering education/practice the idea has not yet caught on widely. Nevertheless, we argue that a modified version of this oath is the most appropriate solution to the ethics dilemma in intelligent machine development. But what is essential and what aspects should be modified to improve the oath? What are the right ethics? Most probably, there is no universal answer.

It has been pointed out [14] that cultural differences lead to the fact that it is unrealistic and impractical to struggle for an internationally unified code of ethics. There are fundamental differences between eastern and western tradition. In Europe, there is a tendency towards deontology and skepticism as regards to robots, whereas in the USA, the starting point is utilitarian ethics and the fundamental question is then "will robots make 'us' more happy?". And in eastern tradition, i.e. Buddhism, robots are seen as a partner in the global interaction of things. [15]

Hence, commitment to specific ethics will be subject to cultural influences. It is likely that there will be a large number of engineering ethics practiced in different places or within different organizations, including working without specific ethics at all. We consider this lack of universality and cultural ambiguity the major strength of the approach. Even if this oath is something like Luther's small catechism and people cherry-pick of it what they see fit it nevertheless creates awareness for the ethical dimensions of engineering and brings transparency to underlying design decisions.

4. SYSTEM DESIGN AND OATH

To put the concept of the oath into effect, we present our view on how the oath can be best implemented in the development of an autonomous robot.

An autonomous robot belongs to the category of autonomous machines and is an intelligent machine. The machines operate not only in the physical realm of forces and motions but also in the abstract realm of information [16]. As a result there is a high level of complexity in intelligent machines. The autonomous robot is a collection of different components, attributes, and relationships that together produces results. Thus the robot can be defined as a complex system and as it combines structural components with activities it is further classified as a dynamic system [17]. As the robot interacts with its environment and allows information, energy, and matter to cross its boundaries it is further classified as an open system. Therefore, the autonomous robot is a complex technical system and needs a systems engineering design approach to enable a successful realization. In this section, technical system will be used synonymously to mean autonomous robot.

Systems engineering (SE) is not a traditional engineering in the same sense as electrical engineering, mechanical engineering or any of the other engineering specialties. One definition of SE is, “an approach to translate operational needs and requirements into operationally suitable blocks of systems” [18]. The approach consists of a top down iterative process of requirement analysis up to design synthesis, verification, system analysis, and control. SE principles influence the balance between performance, risk, cost, and schedule of a system design. The main task of the engineers is to apply their engineering and scientific knowledge to the solution of technical problems.

The critique of Asimov´s laws reveals that a technical system needs two capabilities: responsiveness and smooth transfer of control [2]. These laws are system-centric and disregard the role of the designers. The mental modeling and creation of a technical system in SE is the task of design and development engineers, referred here as designers. The designers carry a heavy burden of responsibility, since their ideas, knowledge and skills determine the type of technical system. Thus the designer’s activities affect almost all areas of human life and are responsible for the behavior of the technical system. Murphy and Woods [2] proposed three alternative laws aimed at designers. The alternative three laws emphasize responsibility and resilience, concentrating on safety oriented designs, responsiveness, and smooth transfer of control. The three laws are all system requirements addressing the technical issues in a systems engineering process [18]. These laws fall short of having a profound effect on design practices as it serves only as a reminder for the designer of his legal and professional responsibility. These laws will allow the designer to avoid responsibility difficult to implement in a real time situation.

To address the difficulties of applying Asimov’s and the alternative laws, we propose the alternative oath concept to be considered and included in an engineering design process. There is numerous system considerations that should be identified and studied when developing design criteria, many of these are shown in Fig. 1 [18].

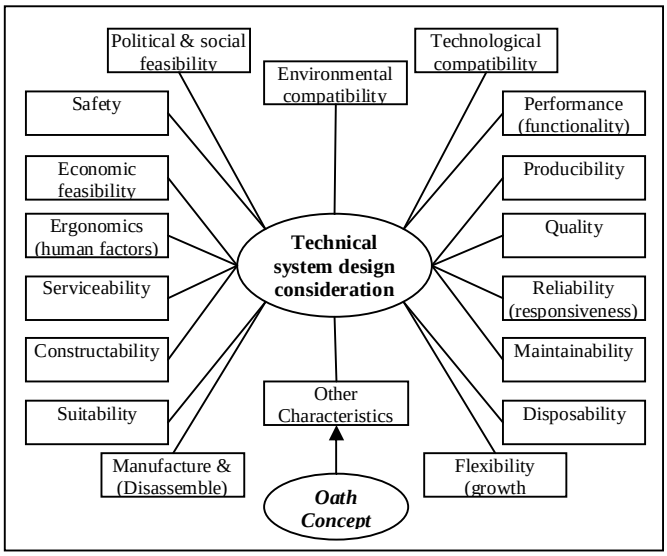

Figure 1. Identification of system design consideration

Design considerations provide a broad range of possibilities from which the derivation of the design criteria may proceed. An essential design activity within system design consideration process should include the oath concept “Fig. 1”. The oath concept must be inherent within the system engineering process and must be invoked regularly as the system design activity progresses.

We consider the oath concept to be implemented right at a conceptual design stage in the design process as an engineering requirement. Requirements are the input for design and operational criteria, and criteria are the basis for the evaluation of technical system safety, performance, responsiveness, quality, and reliable configuration [17], [18], [20], [21]. Conceptual design phase is part of the design process, which begins with a design need and requirements and is preceded by a design decision [19]. This is a critical stage for the designer as the essential design problem is identified through abstraction, studies, and mental modeling. Thus the oath concept as a

requirement will also be thoroughly studied an oath taking before a decision is made to continue to the next phase in the design process, Fig. 2.

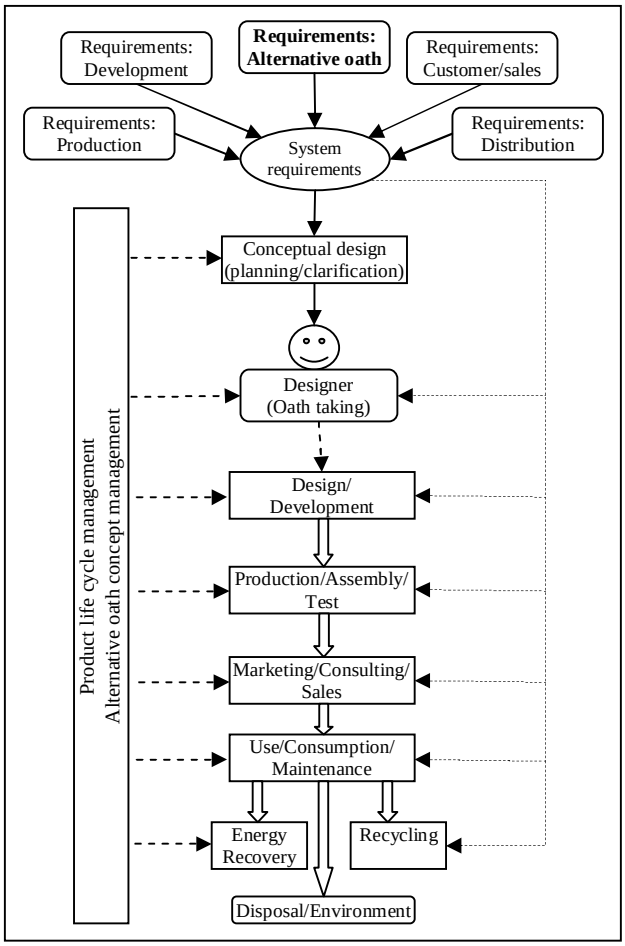

Figure 2. Technical system life cycle design including alternative oath requirements

5. Conclusion

Over the past years, the question of ethics in the context of autonomous machines has changed from being a topic of science fiction to a question in scholarly debates to an urgent practical matter. The massive deployment of military robots has diminished the hopes that something as simple as Asimov's three laws could solve this problem. This has prompted various authors to suggest their own laws of robotics. Interestingly, most of them are both vague and not meant to be implemented within the robot. A good example is the first of the three alternative laws devised by Murphy and Woods: "a human may not deploy a robot without the human-robot work system meeting the highest legal and professional standards of safety and ethics". This is, as the authors admit, not the attempt to create a law for the robot but a demand for guidelines to robot engineering.

Due to cultural differences it would be challenging to struggle for waterproof common ground for roboethics. While it is true that the most appropriate uses for robots might be dependent on cultural factors, we disagree with the notion that it is impossible to find some common ground on ethical issues. There are several other instances where almost universally accepted rules have been developed, most prominently United Nations declaration of human rights. We feel that the initiative with respect to ethics cannot be left to the market. Governments and international organizations should promote a generally accepted code of ethics for intelligent machine developers. This would help the robotics industry to gain more trust from the general public. We are of the opinion that a global codification of roboethics is needed. This codification should include an intercultural subset of ethical subroutines implementable globally by moral artificial agents (AMAs). The proposed solution should especially address the litmus test of killer applications. Creating robots that are designed to harm humans is a major ethical dilemma.

We argue that engineering ethics codified in a developer's oath is a suitable form of implementing the requested compliance with professional standards. We feel the oath provides the appropriate level of latitude to approach the problems immediately at hand. It increases developers' awareness of ethical machine development and gives the freedom of action to make individual decisions. Moreover, good working practices are essential to reducing residual risk related to malfunction in autonomous machines. The development of culturally adapted oaths can be seen as a mean to achieve more diversity in machine development. This diversity in turn is likely to broaden horizons and to increase general ethical awareness.

Acknowledgment

This article is based on a group work on a research course "*BitBang – rays to the future*". We would like to thank Professor Yrjö Neuvo and all other organizers of the course. We show our gratitude to all guest lecturers as well as fellow course students for insightful thoughts and discussions. Also, we are grateful for support provided by the Multidisciplinary Institute of Digitalisation and Energy (MIDE), Aalto University School of Science and Technology.

References

[1] N.E. Sharkey, "The ethical frontiers of robotics", *Science*, vol. 322, no. 5909, pp. 1800-1801, Dec. 2008, DOI: 10.1126/science.1164582.

[2] R. Murphy, D. Woods, "Beyond Asimov: The three laws of responsible robotics", *Intelligent systems*, vol. 24, no. 4, pp. 14-20, July/Aug. 2009.

[3] R. Arkin, "Governing lethal behavior: embedding ethics in a hybrid deliberative/reactive robot architecture", *ACM/IEEE international conference on Human robot interaction*, pp. 121-128, March 12-15, 2008.

[4] R. Sparrow, "Killer Robots", *Journal of Applied Philosophy*, vol. 24, no. 1, pp. 66-77, Feb. 2007.

[5] J. H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics", *IEEE Intelligent systems*, vol. 21, pp. 18-21, July-Aug. 2006.

[6] D. Johnson, "Computer systems: Moral entities but not moral agents", *Ethics and Information Technology*, vol. 8, no. 4, pp. 194-204, Nov. 2006.

[7] J. Haidt, "The new synthesis in moral psychology", *Science*, vol. 316, no. 5827, pp. 998-1002, May 2007.

[8] In Klein v. U.S. (13 Av.Cas. 18137 [D.Md. 1975])

[9] Wells v. U.S. (16 Av.Cas. 17914 [W.D.Wash. 1981])

[10] R. Freitas Jr., "The Legal Rights of Robots", http://www.rfreitas.com/ Astro/LegalRightsOfRobots.htm, Student Lawyer 13, January 1985, pp. 54-56.

[11] A. Siitonen, "Insinöörin etiikka" in Airaksinen, T. (ed.): Ammattien ja ansaitsemisen etiikka, Helsinki: Yliopistopaino, 1991.

[12] C. Allen, W. Wallach, I. Smith, "Why Machine Ethics?", IEEE Intelligent Systems, vol. 21, no. 4, pp. 12-17, July/Aug. 2006.

[13] Towards a Universal Ethical Oath for Scientists, http://portal.unesco.org/shs/en/files/6500/10951486621Ethical_oath_sci.pdf/Ethical_oath_sci.pdf

[14] S. Guo and G. Zhang, "Robot Rights," Science, vol. 322, February 13, 2009, pp. 876

[15] R. Capurro, "Ethics and Robotics: An Intercultural Perspective", Symposium Ethics and Robotics, University of Tsukuba Japan. www.capurro.de/roboethics_japan09.ppt October 3, 2009.

[16] Alanen et al., "Smart Machines and Systems. Recent Advances in Mechatronics in Finland," *Helsinki University of Technology Publications in Machine Design*, Jan. 2001.

[17] E. Rechtin, "Systems Architecting of Organizations: Why Eagles Can't Swim (Systems Engineering)," *CRC Press*, ISBN-10: 0849381401, 2000.

[18] B. Blanchard, W. Fabrycky, *Systems engineering and analysis 4th edition*, Pearson Prentice Hall, ISBN 0-13-196326-0, 2006, pp. 2-21.

[19] G. Pah, W. Beitz, *Engineering Designs: a Systematic Approach*. *Third Edition*, Springer-Verlag. London, 2007, pp. 14-25.

[20] K. Otto, K. Wood, *Product Design: Techniques in Reverse Engineering and New Product Development*. Prentice Hall, Upper Saddle River, NJ, 2001.

[21] G. Altshuller, *Creativity as an exact science*. Gordon & Breach, Luxembourg, 1984.