Sharing Moral Responsibility with Robots: A Pragmatic Approach

Gordana DODIG-CRNKOVIC[1] and Daniel PERSSON[2]

Department of Computer Science and Electronics, Mälardalen University, Västerås, Sweden

Abstract. The increased use of autonomous, learning intelligent systems is causing a new division of tasks between humans and technological artifacts, forcing us to examine the related distribution of responsibilities for their design, production, implementation and use. The focus of our analysis is on the moral responsibility in learning intelligent systems and producer-user relationship mediated by intelligent adaptive and autonomous technology. An argument is made that intelligent artifact can be ascribed "responsibility" in the same pragmatic (functionalist, instrumental) way as they are ascribed "intelligence". We claim that having a system which "takes care" of certain tasks intelligently, learning from experience and making autonomous decisions gives us reasons to talk about a system (an artifact) as being "responsible for a task". No doubt, technology is morally significant for humans, so the "responsibility for a task" with moral consequences could be seen as moral responsibility. However, technological artifacts at the present stage of development can hardly be ascribed any significant degree of moral autonomy. Technological artifacts are products of human design, and shaped by our values and norms. They can be seen as a part of a socio-technological system with distributed responsibilities, similar to safety critical technologies such as e.g. nuclear power or transports. Primarily producers (responsible for the design and manufacturing/programming of intelligent systems) should be attributed responsibility for the consequences of their deployment. Knowing that all possible abnormal conditions of a system operation can never be predicted, and no system can ever be tested for all possible situations of its use, the responsibility of a producer is to assure proper functioning of a system under reasonably foreseeable circumstances. Additional safety measures must however be in place when accidents happen in order to mitigate their consequences. Especially for the autonomous intelligent and learning systems, a number of safety barriers must be in place to prevent unwanted, possibly disastrous effects. The socio-technological system aimed at assuring a beneficial deployment of intelligent systems has several functional responsibility feedback loops which must function properly: the awareness and procedures for handling of risks and responsibilities on the side of designers, producers, implementers and maintenance personnel as well as the understanding of society at large of the values and dangers of intelligent technology. The basic precondition for developing of this socio-technological control system is education of engineers in ethics and keeping alive the democratic debate on the preferences about future society.

Keywords. Intelligent agents, Moral responsibility, Safety critical systems

[1] gordana.dodig-crnkovic@mdh.se

[2] dpn04001@student.mdh.se

Introduction

Engineering can be seen as a long-term, large-scale social experiment since the design, production and employment of engineered artifacts can be expected to have long-range effects [1]. Especially interesting consequences might be anticipated if the engineered artifacts are intelligent, adaptive and autonomous, such as robotic systems. Recently, Roboethics, a field of applied ethics, has developed with many interesting novel insights; see [2]. Ethical challenges addressed within Roboethics include the use of robots, ubiquitous sensing systems and ambient intelligence, direct neural interfaces and invasive nano-devices, intelligent soft bots, robots aimed at warfare, and similar which actualize ethical issues of values (what sort of society is a good one or a desirable one), responsibility, liability, accountability, control, privacy, self, (human) rights, and similar [3].

If the prediction of Veruggio comes true, and the "internet of things" becomes reality "… the net will be not only a network of computers, but of robots, and it will have eyes, ears and hands, it will be itself a robot" [2]. This envisaged robot-net will indeed present ethical challenges never seen before. In accordance with the precautionary principle [4], we have not only rights but also moral obligation to elucidate the ethical consequences of the possible paths of development. Here are some relevant voices of concern[3]: Crutzen [5] wonders if humans are in danger to become just objects of artificial intelligence. Becker discusses the construction of embodied conversational agents which presents a new challenge in the field of both cognitive AI and human-computer-interface. One of the aims of constructing intelligent artifacts is gaining new insights in cognition and communication. At the same time the existence of intelligent artifacts changes our capabilities and behaviors, so we change the technology which in its turn changes us [6]. Consequently: "In shaping responsibility ascription policies one has to take into account the fact that robots and softbots - by combining learning with autonomy, pro-activity, reasoning, and planning - can enter cognitive interactions that human beings have not experienced with any other non-human system" [7].

An interesting question arises as to what happens when cognitive capabilities of autonomous intelligent agents surpass those of humans. Are we going to have any means to control such systems? Are we going to be able to build in an "off button" to save us in case a cognitively superior intelligent system starts to behave inappropriately? It is good to address those questions while we are developing new intelligent and autonomous learning technologies. Riegler suggests a "domestication" strategy for the relation between humans and cognitive robots, in analogy with our taming of wild animals that are also intelligent cognitive agents with whom we have learned to collaborate even though some of them are stronger, faster or in other respects superior to humans [8]. This means gradual mutual adaptation. However, increasingly superior intelligence might imply unique challenges no animal has ever presented to humanity.

[3] Vol. 6, IRIE, 2006 *International Review of Information Ethics* was dedicated to Ethics of Robotics, see http://www.i-r-i-e.net/archive.htm

1. Autonomous, adaptive intelligent systems

We expect intelligent systems to always behave rationally: they should make as good decisions as possible, taking into account current percepts, built-in knowledge and capabilities given beforehand as well as previous experiences acquired by learning through interaction with the environment [9]. The learning capabilities together with other intelligent and adaptive features of an intelligent agent are contained in its software (and sometimes also implemented in the hardware).

For the learning computer system there is no explicit formulation by the producer how the system will perform a certain task. A learning system generalizes knowledge based on a limited set of inputs. This means that it cannot be expected to optimally handle all possible instances of a certain task. These two features are the primary causes of uncertainty in ascribing moral responsibility to a producer in regard to decisions made by learning systems. A lack of control and occasional sub-optimal behavior can be seen as inherent features [10]. Sub-optimal behavior, however, is not something new that learning introduces; errors and bugs have always been an unavoidable part of computer systems [1].

2. Moral responsibility

2.1. Classical approach to moral responsibility, causality and free will

A common approach to the question of moral responsibility is presented in Stanford Encyclopedia of Philosophy, according to which "A person who is a morally responsible agent is not merely a person who is able to do moral right or wrong. Beyond this, she is accountable for her morally significant conduct. Hence, she is an apt target of moral praise or blame, as well as reward or punishment." [12]

In order to decide whether an agent is morally responsible for an action, it is believed to be necessary to consider two parts of the action: causal responsibility and mental state [11]. In this view the mental state aspect of a moral action is what distinguishes morally responsible agents from just causally responsible agents. Traditionally only humans are considered to be capable of moral agency. The basis of the human capability of action is intention, and this internal state is seen as the origin of an act that, depending on the effects it causes, can imply moral responsibility [13][14]. Moreover, intentionality enables learning from mistakes, regret of wrongs and wish to do right – all of which are eminently human abilities. Both causal responsibility and intentionality are necessary for someone to be considered morally responsible for an act.

2.2. Pragmatic (functional) approach to moral responsibility

However, questions of intentionality and free will of an agent are difficult to address in practical engineering circumstances, such as development and use of intelligent adaptive robots. Dennett and Strawson suggest that we should understand moral responsibility not as individual duty but instead as *a role defined by externalist pragmatic norms of a group* [15][16]. We will also adopt a pragmatic approach, closer to actual robotic applications, where the question of free will is not the main concern. Moral responsibility can best be seen as a social regulatory mechanism which aims at

enhancing actions considered to be good, simultaneously minimizing what is considered to be bad.

Responsibility of intelligent systems with different degrees or aspects of intelligence can be compared with the responsibility of infants or highly intelligent animals. Legally it is the parent of a child or the owner of an animal who is responsible for their possible misbehavior. Responsibility in a complex socio-technological system is usually a distribution of duties in a hierarchical manner, such as found in military organizations. Dennett views moral responsibility as rational and socially efficient policy and as the result of natural selection within cooperative systems [15]. Similarly, "Since the aim of morals is to regulate social cooperation, we should view moral rules as principles that make cooperation possible if the majority of the members of society follow these", [17]. In this functionalist view of responsibility, what matters is the regulative role responsibility plays for the behavior of an agent within a socio-technological system. Moral responsibility as a regulative mechanism shall not only locate the blame but more importantly assure future appropriate behavior of a system.

In a pragmatic spirit moral responsibility is considered to be the obligation to behave in accordance with an accepted ethical code [18]. Moral responsibility matters as long as it influences the behavior of individuals who have been assigned professional responsibilities [19]. In e.g. Software Engineering practice moral responsibility is seen as a subfield of system dependability. The practical questions of allocation, acceptance, recording and discharge of responsibilities and their reflection in that context are addressed in the DIRC project [20]. The general problem of practice is how to draw boundaries around domains of distinct responsibilities, the scope of differing sets of ethical principles, the limits of various levels of trust, and how to reflect these boundaries in an information system.

When attributing moral responsibility, the focus is usually on individual moral agents. Coleman and Silver argue that even corporations and similar groups in socio-technological systems also have a *collective moral responsibility* [21] [22].

3. Moral responsibility of intelligent systems

A common argument against ascribing moral responsibility to artificial intelligent systems is that they do not have the capacity for mental states like intentionality, and thus cannot fulfill all requirements for being morally responsible [13][14]. The weakness of this argument is that it is actually nearly impossible to know what such a mental state entails, and to gain the kind of access to an agent's mental state that would be required to assert whether states of intention are present or not [23]. In fact, even for humans, intentionality is ascribed based on the *observed behavior* as we have no access to the inner workings of human minds – much less than we have access to the inner workings of computing systems.

Another argument against ascribing moral responsibility to artificial intelligent systems maintains that it is pointless to assign praise or blame to an agent of this type when it has no meaning to the agent. The only rational way of dealing with such agents is to physically alter them, either by correcting or removing them [23].

Both above arguments stem from a view in which artificial intelligent systems are seen primarily as isolated entities. We argue that to address the question of ascribing moral responsibility to intelligent systems we must see them as parts of larger socio-technological systems. From such a standpoint ascribing responsibility to an intelligent

system has primarily a regulatory role. Investigation of moral responsibility for actions involving any technological artifacts must take into account actions of the users and producers of the artifacts, but also the technological artifacts themselves [24]. It is not only human agents that by engineering and operating instructions can influence the morality of artificial agents. Artifacts, as actors in socio-technological systems, can impose limits on and influence the morality of human actors too [25][26]. Despite this, the existence and influence of artifacts up to now have always originated in a human designer and producer [24]. Delegating a task to a machine is also delegating responsibility for the safe and successful completion of that task to the machine [24]. Delegation of tasks is followed by distribution of responsibilities in the socio-technological system, and it is important to be aware of how they influence the balance of responsibilities between different actors in the system [25]. Commonly the distribution of responsibility for the production and use of a system can be seen as a kind of contract, a manual of operations, which specifies how and under what circumstances the artifact or system should be used [10]. The user is responsible for using the artifact or system in accordance to the (producer supplied) specifications of the manual, and the producer is responsible for assuring that the artifact or system really functions as specified [10]. This clear distinction between the responsibilities of producer and user was historically useful, but with increased distribution of responsibilities throughout a socio-technological system this distinction becomes less clear. Production and use of intelligent systems has increased the difficulty, as the intelligent artifacts themselves display autonomous morally significant behavior, which has lead to a discussion (see [10][13][23][27]) about the possibility of ascribing moral responsibility to machines. Many of the practical issues in determining responsibility for decisions and actions made by intelligent systems will probably follow already existing models that are now regulated by product liability laws [28]. There is a suspicion though that this approach may not be enough and that alternative ways of looking at responsibility in the production and use of intelligent systems may be needed [29].

In sum, we claim that having a system which "takes care" of certain tasks intelligently, learning from experience and making autonomous decisions gives us good reasons to talk about a system as being "responsible for a task". No doubt, technology is morally significant for humans, so the "responsibility for a task" with moral consequences could be seen as moral responsibility. The *consequential responsibility* which presupposes *moral autonomy* will however be distributed through the system.

Numerous interesting questions arise when the issue of artificial agents capable of being morally responsible in the classical sense is addressed by defining autonomous ethical rules of their behavior – questions addressed within the field of Machine Ethics [30]. Building in ethical rules of behavior for e.g. softbots seems to be both useful and practical.

4. Distribution of responsibilities and handling of risks in intelligent technology

Based on the experiences with safety critical systems such as nuclear power, aerospace and transportation systems one can say that the socio-technological structure which supports their beneficial functioning is a system of safety barriers preventing and mitigating malfunction. The central and most important part is to assure the safe

functioning of the system under normal conditions, which is complemented by the supposed abnormal/accidental condition scenarios. There must be several levels of organizational and physical barriers in order to cope with different levels of severity of malfunctions [31].

As already pointed out, one of the concerns of design and production is uncertainty. There are uncertainties in every design process that are the result of our limited resources. All new products are tested under certain conditions in a given context. Computer programs for example are tested for a large but finite number of cases, and the "rest-risk" is assessed based on best practices [35]. This implies that an engineered product may, sooner or later, in its application be used under conditions for which it has never been tested. Even in such situations we expect the product to function safely. Handling risk and uncertainty in the production of a safety critical technical system is done on several levels. Producers must take into account everything from technical issues, through issues of management and of anticipating use and effects, to larger issues on the level of societal impact [32]. The central ethical concerns for engineers are: "How to evaluate technologies in the face of uncertainty" and "How safe is safe enough" [33]. Risk assessment is a standard way of dealing with risks in the design and production of safety critical systems [34], also relevant for intelligent systems.

In the discussion of technology affecting the environment there is an increased emphasis on what is known as *the precautionary principle*, which consists of four parts: preventing harm, laying the burden of proof of harmlessness on proponents of new technology, examining alternatives, and making decisions about technology in a democratic way [4]. Any technology subject to uncertainty and with a potentially high impact on human society is expected to be handled cautiously, and intelligent systems surely fall into this category. Thus, preventing harm and having the burden of proof of harmlessness is something that producers of intelligent systems are responsible for.

A precondition for this socio-technological control system is an engineer informed about the ethical aspects of engineering, where education in professional ethics for engineers is a fundamental factor [19].

5. Conclusion

According to the classical approach, free will is essential for an agent to be assigned moral responsibility. Pragmatic approaches (Dennett, Strawson) on the other hand focus on social, organizational and role-assignment aspects of responsibility. We argue that moral responsibility in intelligent systems is best viewed as a regulatory mechanism, and follow essentially a pragmatic (instrumental, functionalist) line of thought. We claim that for all practical purposes, the question of responsibility in safety critical intelligent systems may be addressed in the same way as the safety in traditional safety critical systems such as nuclear industry and transports. As a practical example of application we can mention that in Software Engineering responsibility is addressed as a part of software dependability of a software system.

Long-term, wide range consequences of the deployment of intelligent systems in human societies must be discussed on a democratic basis as the intelligent systems have a potential of radically transforming the future of humanity. Education in professional ethics for engineers is a fundamental factor for building a socio-technological system of responsibility.

References

[1] Martin, M.W., Schinzinger, R., *Ethics in Engineering*, McGraw-Hill, 1996.

[2] Roboethics: http://www.roboethics.org, http://www.scuoladirobotica.it, http://roboethics.stanford.edu/; http://ethicalife.dynalias.org/schedule.html http://www-arts.sssup.it/IEEE_TC_RoboEthics/; http://ethicbots.na.infn.it/ http://www.capurro.de/lehre_ethicbots.htm ETHICBOTS seminar

[3] Dodig-Crnkovic G., Professional Ethics in Computing and Intelligent Systems, *Proceedings of the Ninth Scandinavian Conference on Artificial Intelligence* (SCAI 2006), Espoo, Finland, October 25-27, 2006.

[4] Montague P. , The Precautionary Principle, *Rachel's Environment and Health Weekly*, No. 586, 1998, Available: http://www.biotech-info.net/rachels_586.html

[5] Crutzen C.K.M., Invisibility and the Meaning of Ambient Intelligence, *International Review of Information Ethics*, Vol. 6, IRIE, 2006, pp. 52-62.

[6] Becker B., Social Robots - Emotional Agents: Some Remarks on Naturalizing Man-Machine Interaction, *International Review of Information Ethics*, Vol. 6, IRIE, 2006, pp. 37-45.

[7] Marino D. and Tamburrini G., Learning Robots and Human Responsibility, *International Review of Information Ethics*, Vol. 6, IRIE, 2006, pp. 46-51.

[8] Riegler A., The Paradox Of Autonomy: The Interaction Between Humans And Autonomous Cognitive Artifacts, in Dodig-Crnkovic G. and Stuart S., eds., *Computation, Information, Cognition – The Nexus and The Liminal*, Cambridge Scholars Publishing, Cambridge, 2007.

[9] Russell S. and Norvig P., *Artificial Intelligence – A Modern Approach*, Pearson Education, Upper Saddle River, NJ, 2003.

[10] Matthias A., The responsibility gap: Ascribing responsibility for the actions of learning automata, *Ethics and Information Technology*, Vol. 6, Kluwer Academic Publishers, 2004, pp. 175-183.

[11] Nissenbaum, H.,"Computing and Accountability", *Communications of the ACM*, Vol. 37, ACM, 1994, pp. 73-80.

[12] Eshleman A., Moral Responsibility, *The Stanford Encyclopedia of Philosophy (Fall 2004 Edition)*, Edward N. Zalta (ed.), Available: http://plato.stanford.edu/archives/fall2004/entries/moral-responsibility/

[13] Johnson D. G., Computer systems: Moral entities but not moral agents, *Ethics and Information Technology*, Vol. 8, Springer, 2006, pp. 195-204.

[14] Johnson D. G. and Miller K. W., A dialogue on responsibility, moral agency, and IT systems, *Proceedings of the 2006 ACM symposium on Applied computing table of content*, Dijon, France, 2006, pp. 272 – 276.

[15] Dennett, D. C., Mechanism and Responsibility, in *Essays on Freedom of Action*, T. Honderich (ed), Routledge & Keegan Paul, Boston, 1973.

[16] Strawson P. F., Freedom and Resentment, in *Freedom and Resentment and Other Essays*, Methuen, 1974.

[17] Järvik M., How to Understand Moral Responsibility?, *Trames*, No. 3, Teaduste Akadeemia Kirjastus, 2003, pp. 147–163.

[18] Sommerville, I., http://www.comp.lancs.ac.uk/computing/research/cseg/projects/dirc/Responsibility/

[19] Dodig-Crnkovic G., On the Importance of Teaching Professional Ethics to Computer Science Students, Computing and Philosophy Conference, E-CAP 2004, Pavia, Italy, in L. Magnani, ed., *Computing and Philosophy*, Associated International Academic Publishers, 2005.

[20] DIRC project: http://www.comp.lancs.ac.uk/computing/research/cseg/projects/dirc/projectthemes.htm

[21] Coleman, K. G., Computing and Moral Responsibility, *The Stanford Encyclopedia of Philosophy (Spring 2005 Edition)*, Edward N. Zalta (ed.), Available: http://plato.stanford.edu/archives/spr2005/entries/computing-responsibility/

[22] Silver D. A., Strawsonian Defense of Corporate Moral Responsibility, *American Philosophical Quarterly*, Vol. 42, 2005, pp. 279-295.

[23] Floridi L. and Sanders J. W., On the morality of artificial agents, *Minds and Machines*, Vol. 14, Kluwer Academic Publishers, 2004, pp. 349-379.

[24] Johnson D. G. and Powers T. M., Computer systems and responsibility: A normative look at technological complexity, *Ethics and Information Technology*, Vol. 7, Springer, 2005, pp. 99-107.

[25] Adam A., Delegating and Distributing Morality: Can We Inscribe Privacy Protection in a Machine?, *Ethics and Information Technology*, Vol. 7, 2005, pp. 233-242

[26] Latour B., Where are the missing masses, sociology of a few mundane artefacts, originally in Wiebe Bijker and John Law, eds., *Shaping Technology-Building Society. Studies in Sociotechnical Change*, MIT Press, Cambridge Mass. pp. 225-259, Available: http://www.bruno-latour.fr/articles/1992.html.

[27] Stahl, B.C., Information, Ethics, and Computers: The Problem of Autonomous Moral Agents, *Minds and Machines*, Vol. 14, Kluwer Academic Publishers, 2004, pp. 67-83.
[28] Asaro P. M., Robots and Responsibility from a Legal Perspective, *Proceedings of the IEEE 2007 International Conference on Robotics and Automation, Workshop on RoboEthics*, Rome, April 14, 2007
[29] Stahl B. C., Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency, *Ethics and Information Technology*, Vol. 8, Springer, 2006, pp. 205-213.
[30] Moor, J. H., The Nature, Importance, and Difficulty of Machine Ethics, *IEEE Intelligent Systems*, IEEE Computer Society, 2006, pp. 18-21.
[31] Dodig-Crnkovic G., ABB Atom's Criticality Safety Handbook, ICNC'99 Sixth International Conference on Nuclear Criticality Safety, Versailles, France, (1999) , Available: http://www.idt.mdh.se/personal/gdc/work/csh.pdf
[32] Huff, C., Unintentional Power in the Design of Computing Systems, in T. W. Bynum and S. Rogerson, eds., *Computer Ethics and Professional Responsibility*, Blackwell Publishing, Kundli, India, 2004, pp. 98-106.
[33] Shrader-Frechette K., Technology and Ethics, in R. C. Scharff and V. Dusek, eds., *Philosophy of Technology - The Technological Condition*, Blackwell Publishing, Padstow, United Kingdom, 2003, pp. 187-190.
[34] Stamatelatos M., Probabilistic Risk Assessment: What Is It And Why Is It Worth Performing It?, NASA Office of Safety and Mission Assurance, 2000, Available: http://www.hq.nasa.gov/office/codeq/qnews/pra.pdf
[35] Larsson M., Predicting Quality Attributes in Com-ponent-based Software Systems, PhD Thesis, Mälardalen University Press, Sweden, ISBN: 91-88834-33-6, 2004