Can Humans Be Replaced by Autonomous Robots?

Ethical Reflections in the Framework of an Interdisciplinary Technology Assessment

Michael Decker, ITAS – Institute for Technology Assessment and Systems Analysis, Research Centre Karlsruhe

***Abstract* – Autonomous robots capable of learning are being developed to make it easier for human actors to achieve their goals. As such, robots are primarily a means to an end and replace human actions (or parts of them). An interdisciplinary technology assessment was carried out to determine the extent to which a replacement of this kind makes sense in terms of technology, economics, legal aspects and ethics. With reference to learning robots, the focus was particularly on the central question of whether learning robots represent a grey area of liability between manufacturer and owner. Proceeding from an ethical perspective in this article, the proposal is made – derived from Kant's formula of humanity – that robot learning should be anchored in the responsibility of the robot's owner.**

I. Introduction

Robots are one of those rare technical systems whose potential in terms of construction and impact were comprehensively described and discussed before they were actually built. However, if one looks at the developments made over the past 10 years, where significantly more microcontrollers are in use *outside* computers (e.g., microprocessors in cars, aeroplanes, houses, machine controls, satellites, mobile telephones, washing machines, gaming machines, and cameras), then one can speak of a robotisation of the human environment. This development goes hand in hand with greater power, miniaturisation and broad availability in other areas: engines, drives, batteries, materials, connections, and sensors. General robotics has now achieved a status such that working with them no longer necessarily entails an odyssey with countless forced interruptions through such diverse fields as mechanical engineering, electrical engineering, control engineering, software engineering and algorithmics. On the contrary, one can assume that the hardware and software components exhibit a (minimal degree of) availability and stability. Systems can be constructed systematically to engineering standards.

Manuscript received January 18, 2007.

M.K. Decker is with the Institute for Technology Assessment and Systems Analysis at the Forschungszentrum Karlsruhe.
Hermann-von-Helmholtz-Platz 1, D-76344 Eggenstein-Leopoldshafen, Germany. Phone: ++49-(0)7247-82-3007; fax: ++49-(0)7247-82-4806; email: decker@itas.fzk.de

These developments in robotics research permit the use of robots in types of applications not previously touched by technology. Robots then act in contexts in which humans previously acted. In a sense, in these contexts humans are being replaced by robots. The question of whether humans can be replaced in specific contexts of action is formulated very generally and was a central issue in an interdisciplinary technology assessment [1].

The first issue to be addressed is *technical replaceability*. A robot will only be used if it is technically capable of carrying out actions which are necessary to fulfil a specific task. A robot will always be technically evaluated in a means-end context. It competes with other means that may be available and that could also be used to achieve this end. The technical criteria for the use of robots are derived from this means-end context.

With means-end contexts of this kind, one swiftly reaches a point at which the utility of robots must be assessed (*economic replaceability*). This is not merely a case of strict cost-benefit considerations in the sense of economics, since these often neglect certain aspects of the concept of utility used in a broader sense. This becomes obvious in evaluating so-called "service robots" since "service" is always associated with qualities that are hard to assess such as friendliness, helpfulness, attentiveness, and politeness. How can these "soft skills" be included in the evaluation? How can one justify the right of the demand side to have these additional service aspects?

The position of the demand side – the customers – must also be discussed from a different perspective. It is, for instance, necessary to examine whether the introduction of robots as actors leads to any changes from a legal viewpoint (*legal replaceability*). On the one hand, liability issues are certainly relevant here. Who is liable for damage caused by a robot? On the other hand, questions regarding consumer protection arise, since in future scenarios robots and people with only a layman's understanding of robots will come into contact with each other. Do robots have to be especially equipped for these "unexpected" encounters? Or must humans be specially trained to deal with such potential situations? Are there any additional aspects to be considered when dealing with "learning" robots?

Finally, we must address the question of whether there are action contexts in which a modern society should exclude the integration of robots. These could include care of the sick

and elderly, for example, or the education of children. In evaluating the question of the areas in which autonomous robots should act instead of humans, we are addressing the question of *ethical replaceability*. In the above-mentioned means-end context, such ethical reflection is aimed particularly at the ends level [2]. It is thus a question of whether the ends involved in the use of autonomous robots are ethically justifiable.

In the following, we discuss how robot systems with modern learning algorithms, which take the experimental nature of learning into account, should be evaluated in this „structure of replaceability". Modern learning algorithms are evaluated on the basis of results from Christaller et al. [1], in which the liability for robots is discussed. In particular, the question arises as to whether difficulties in technical realisation will not doom the recommendations for action agreed upon at that time to failure.

II. "Autonomous" Robots and Kant's Formula of Humanity

When applied to cooperation between humans and technical artefacts, Kant's formula of humanity says that the human actor in this cooperation may not be instrumentalised, i.e., used solely as a means to achieve a particular end. The ethical inadmissibility of instrumentalisation can be derived from Kant's second formulation of the categorical imperative, the formula of humanity, in which he seeks to set the ethical dignity of humans in a moral philosophical context. According to this, it is a violation of human dignity for a person to be made solely a means for the arbitrary use toward an external end.[1] The formula of humanity in the categorical imperative protects the autonomy and dignity of humans.

The formula of humanity is widely accepted in the main currents of ethics today. The absolute validity claimed for it by Kant is, however, disputed by the consequentialist and particularly the utilitarian positions. For these positions, restrictions on the autonomy and dignity of individual persons are accepted under certain conditions if they can be justified by superordinate and more extensive considerations of utility [1]. Furthermore and analogously, the target group of the formula of humanity in utilitarianism is not categorically determined but introduced via interpretations of the concepts of interest and feeling. This, for example, makes it possible for the feelings of animals to be included in the calculation. The applicability of the formula of humanity is thus not limited only to humans.

In the ethics of technology, the primary function of the formula of humanity is to discover situations in which means-end relations are reversed or make themselves independent. But yet even in Kant's classical version of the formula of humanity, the inadmissibility of instrumentalisation is not to be understood to mean that persons in action or institutionalisations must be seen as ends in themselves. Ethically inadmissible is solely for humans to be used exclusively as means to achieve ends which are not in their own interests.

Even in this "weak" reading, the formula of humanity is suitable for deriving criteria to assess technology. From it we can deduce a normative limitation and restriction on technological processes in which humans may not simply be made instruments or the object of mere arbitrariness. If one assumes that technological developments first mainly follow the logic of what is possible, in which from a constructive approach no normative regulations are taken into consideration, then it is possible in the framework of an ethical and political discourse to ask for an evaluation and, if appropriate, a revision of the technological process. In this context the ethical inadmissibility of instrumentalisation is an expression of the fact that moral reasons should take priority over simple considerations of utility. A consequence of this is that the competence to specify and alter ends in the context of concrete action belongs exclusively to humans, those who are the potential subjects of autonomy and evaluations. Phrased differently, if as a result of technical processes an individual is no longer capable of acting as a person in the social sphere, then we can assume that the technical constraints exceed the limits of what is acceptable.

The formula of humanity aims – from an ethical viewpoint – to protect the autonomy of humans. The concept of autonomy originally designates the fact that a person establishes for himself/herself and perhaps also for other people a moral law which determines the relevant rules and plans for life. The concept of autonomy – in its meaning as reasonably determined self-legislation – thus differs extensionally and intensionally from the concepts of self-organisation and self-direction since it does not focus on individual action episodes but determines their rules and laws.

Against the background of the transference of the concept of autonomy to other contexts of meaning, in particular that of artificial intelligence research and robotics, the following systematic differentiation can be proposed [1]:

1) First-level autonomy or *technical autonomy*. First-level autonomy is present in cases of complex automation with technically induced degrees of freedom. The attribute autonomy refers here to the characteristic of a machine to carry out directions and actions within defined areas of motion.
2) Second-level autonomy or *personal autonomy*. Autonomy in its precise sense is used to denote the ability of persons to spontaneously adopt attitudes and carry out actions which are in principle not predictable [3]. Personal autonomy takes place in the form of actions in the sphere of reasons. These do not have to be determined morally or, in a narrower sense, rationally. Life plans in the sense of wishes and interests constitute a typical case of personal autonomy.

[1] Kant 1785, p. 429 (translation M.D.): "Act in such a way that you treat humanity, whether in your own person or in the person of another, always at the same time as an end and never as a means".

3) Third-level autonomy or *ideal autonomy* in the realm of ends. Actions in the sphere of reasons may be the object of moral self-determination in the sense of the categorical imperative[2]. Under conditions of third-level autonomy, the actions of persons are exclusively morally determined. Their actions would fuse under ideal conditions into an integral unity.

When speaking of "autonomous" robots, no differentiation is made as a general rule between these various levels of autonomy. A robot is termed "autonomous" merely on account of its being in a position to carry out certain tasks independently on the basis of its sensors and system attributes. This „independence" of a robot must, however, be distinguished from our everyday experience, according to which a person is attributed with autonomy for using his or her discretionary limits and scope for action on the basis of well-determined reasons. Autonomy thus becomes the capacity of a person to determine their own action in a specific context and to formulate the laws, principles and maxims according to which they want to lead their own life.

In the following we look more closely at the cooperation between an autonomous (level 2) human and an autonomous (level 1) robot. When two humans cooperate, the corresponding action is agreed, for example, against the background of a categorical imperative that everyone should behave accordingly in this specific situation. In cooperation between human and robot, in contrast, the ethical inadmissibility of instrumentalisation on the one hand and the achievement of the corresponding goals of cooperation on the other hand represent the relevant criteria.

III. Autonomous Robots with a Capacity to Learn

If robots are to be put into the position of carrying out actions in specific service sectors and different contexts, then they must be able to adapt themselves in some fashion to different contexts of action if their services are to be helpful, particularly where these contexts are complex. In the process, the robot must be able to learn. This begins with its perception of the environment via sensors, continues with the planning of actions on the basis of these sensory data and leads finally, if one assumes that the robot does not always restart from the beginning, to learning – adaptation to the context of action.

If we consider how humans learn, for instance through parental explanation, copying siblings, or by trying something out, then it becomes obvious that trial and error is an integral part of learning. When a person perceives something that is worth learning, they take the next opportunity to try it out. If the trial is successful, what has been learned is validated; if the trial fails, it is questioned.

In robotics research, the central significance of trial and error is taken into account. The most recent research projects place the experimental [5] or playful [6] nature of learning at the centre of their research strategy. Yet this thought is not a new one [7], [8], and it is also a central aspect of the famous robot system "Cog" developed at MIT, which is designed to learn "like a child" from the people around it [9], [10]. To permit this experimental, playful type of learning, artificial neural networks are used (connectionism). These networks represent an attempt to recreate the functional principles of the human brain. Artificial neurones are combined together in such a way that they can exchange signals with one another. Incoming signals are transferred over a weighting factor to the output signals. The "training" of the artificial neural network then reflects the variations in the weightings [11]. An artificial neural network thus represents a signal input-output unit which does not admit any possible interpretation of its internal processes: "In artificial neural networks, the symbolic representation of information and flow control disappears completely: instead of clear and distinct symbols we have a matrix of synaptic weights, which cannot be interpreted directly anymore" [12]. Matthias notes that the same is true for further learning algorithms and deduces from this that there is a gap as to who is responsible for the actions of learning robots. The implementation of a learning algorithm entails that even the robot manufacturer is no longer in a position to predict the robot's actions if the latter has been in a new context of action for some time and has "learned" there. Thus the robot manufacturer can no longer be held liable for the actions of the robot in the usual way.

IV. Recommendations for Acting and Conclusion

Christaller et al. [1] argued in a very similar vein in the final report on the project "Robots: Options for the Replaceability of Man". Proceeding from Kant's formula of humanity, they first pointed out that in cooperation between „man and robot", man is at the top of the decision-making hierarchy. This results in immediate demands being made on the organization of the man-machine interface. The responsibility gap in learning robots that was diagnosed by Matthias [12] is handled in connection with the liability for damages caused by robots. To be precise, the gap in responsibility arises between the robot manufacturer, who is an expert on robots and who implemented the learning algorithm, and its owner, who uses the robot in a particular context of action and who as a rule is not a robot expert. On the one hand, we have the ethical argument that – even with learning robots – man's role as the determining decider in the cooperation must be guaranteed, and on the other the legal argument that it is equally necessary to guarantee how responsibility is divided between the robot's owner and its manufacturer. This results in a recommendation for action regarding the technical equipping of learning robots:

"Dealing with learning robots

It should be possible to distinguish learning robots from non-learning ones since the liability for damages between

[2]The categorical imperative bases human action firmly in a binding moral law: "act according to the maxim which at the same time can make itself a universal law" [4]

manufacturer and owner is influenced by the employment of learning algorithms.

It is recommended that the learning process be transparent for the robot owner and for third parties. In this connection it can be of assistance to install a non-manipulable black box to continuously document the significant results of the learning process or the sensory impulses." [1]

At first sight, this seems to indeed be a technical solution to the problem of the responsibility gap: The learning processes are made transparent to the robot's owner and are documented in a black box similar to that found in an aeroplane. In concrete terms, the robot owner has to confirm, for example by pressing a button that the learning process has been made transparent. In reality, he would thus confirm that he agrees that the robot is carrying out this learning process. This would place learning in the sphere of responsibility of the robot owner. In this case, the robot manufacturer would only have to refer clearly enough in the instructions to the learning algorithm, to the confirmation procedure, and to the fact that this confirmation is recorded in the black box.

If we look more closely, however, it is clear that this very presentable technical solution for an ethical-legal problem represents a significantly greater technical challenge than the mere integration of a „confirmation" button and a recorder. The high hurdle that is obscured is the fact that the robot must be able to communicate to the robot owner what he suggests be learned. If this communication were to take place, for instance, in text form on the robot's screen, then the robot would have to phrase a text in which he describes an observation, formulate a hypothesis on the basis of this observation, and finally develop a suggested explanation and procedure for action, which would then be learned. Expressed more concisely, the robot would have to develop a well-founded if-then statement – if I perceive x, I do y because z – and pass this on to the robot owner. But precisely if-then statements of this kind, such as are employed in expert systems [11], are not available in the playful and experimental learning algorithms mentioned in the previous section. From the way in which, for example, individual neurons can weight their incoming signals, it is impossible to draw conclusions about actions in concrete contexts of action

Put another way, if a robot could do it, it would have reached the second level of autonomy! Whether and, if so, when this could be the case, is still an open question. There is still time for further interdisciplinary analyses [13] that accompany development, in which ethical reflection, as in the case presented here, should take a significant role.

References

[1] T. Christaller, M. Decker, J.-M. Gilsbach, G. Hirzinger, K. Lauterbach, E. Schweighofer, G. Schweitzer, D. Sturma, *Robotik. Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft*. Berlin, Heidelberg: Springer, 2001

[2] C.F. Gethmann, T. Sander, "Rechtfertigungsdiskurse," in *Ethik in der Technikgestaltung. Praktische Relevanz und Legitimation*, A. Grunwald, S. Saupe (Hrsg.) Berlin: Springer, 1999, pp. 117-151

[3] D.M. MacKay, *Freedom of Action in a Mechanistic Universe*. Cambridge: Cambridge UP, 1967

[4] I. Kant (1785), "Grundlegung zur Metaphysik der Sitten", in *Werke*, Akademieausgabe, Band IV. Berlin, 1968

[5] XPERO *Roboter lernen durch Spielen*. Presseerklärung Fachhochschule Bonn-Rhein-Sieg vom 20. Juni 2006, Sankt Augustin

[6] BCCN *Wissenschaftler im Portrait*: Florentin Wörgötter: Roboter, die spielend lernen. Newsletter des Centers for Computational Neuroscience, vol. 06, 2006, Berlin

[7] E. Mjolsness, D. DeCoste, "Machine Learning for Science: State of the Art and Future Prospects", in *Science*, vol. 293, 2001, p. 2051-2055

[8] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, E. Thelen, "Artificial Intelligence: Autonomous Mental Development by Robots and Animals", in *Science*, vol. 91, 2001, S. 599-600.

[9] R.A. Brooks, "The Cog project", in *Journal of the Robotics Society of Japan*, vol. 15, no. 7, 1997, p. 968-970

[10] R.A. Brooks, L.A. Stein, "Building brains for bodies", in *Autonomous Robots*, vol. 1, no. 1, 1994, p. 7-25

[11] M. Decker, "Perspektiven der Robotik. Überlegungen zur Ersetzbarkeit des Menschen", in *Graue Reihe*, Bd. 8, Europäische Akademie Bad Neuenahr-Ahrweiler, 1997

[12] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata", in *Ethics and Information Technology*, vol. 6, 2004, p. 175-183

[13] M. Decker; A. Grunwald, (2001): "Rational Technology Assessment as Interdisciplinary Research". M. Decker (ed.): Interdisciplinarity in Technology Assessment. Implementation and its Chances and Limits. Springer Berlin, 2001, p. 33-60