# Machine Ethics and Human Ethics: A Critical View

**Francesco Amigoni and Viola Schiaffonati**

Artificial Intelligence and Robotics Laboratory, Dipartimento di Elettronica e Informazione, Politecnico di Milano
Piazza Leonardo da Vinci 32, I-20133, Milano
amigoni@elet.polimi.it; schiaffo@elet.polimi.it

**Abstract**

The rapid changes in the relationship between humans and machines evidence a progressive delegation of some human activities to machines. Among those, there are not only physical activities but also intellectual and social activities. However, we deem that even the most powerful computers should be considered as tools created and used by humans and not totally autonomous and independent subjects. Therefore, we see a theoretical objection to machine ethics as the enterprise of adding an ethical dimension to some machines. Rather than trying to implement ethical behavior in machines, we suggest extending the traditional ethical framework. This extension should include also problems concerned with the behavior of machines towards users and other machines in the perspective of the centrality of humans with respect to machines.

## Our Proposal for Machine Ethics

In the following, we briefly present our position on the topic of machine ethics. Due to the preliminary stage of the following ideas, we do not develop in all details our arguments, but we schematically discuss some points to give the flavor of our approach.

- In recent years we have been assisting to a rapid change in the relationship between humans and machines. The most striking result is a progressively fast delegation of some human tasks to machines. Traditionally, machines have been useful tools for substituting humans in some physical activities. What has emerged in the last fifty years is the possibility to delegate to machines also *intellectual* activities. This has been made possible with the advent of *information machines* (e.g., computers) that embed models of some phenomena belonging to the intelligence of humans and to the intelligent interaction between humans and the external world. Moreover, today these sophisticated machines we interact with can be our delegates for some *social* activities as well. If we consider, for example, software agents in e-commerce, here agents act as humans representatives for purchasing goods on the Web. This goes further physical substitution: agents must be able to reason in order to select the best (rational) possibility. Also, they need social abilities to interact with other agents involved in the application (for example, the sellers) that act on the behalf of other humans.
- However, we believe that even the most powerful computers should be considered as *tools*, instruments in the hands of their makers and users [1]. The carrying on of intellectual activities traditionally performed by humans does not imply an autonomous role for computers, but just a delegation of some of these activities to artifacts that are designed and built by humans themselves who, thus, maintain a central role. Not only humans design and build these instruments for their purposes, but also they decide what machines should do. This emphasizes the human political superiority towards information machines and stresses the dependence of computers from human will [4].
- Nowadays, in advanced research in computer science, *autonomy* is largely emphasized as the ability required for carrying on successful activities. Human beings can be progressively substituted only by artificial systems able to act in an autonomous way. However, we believe that the particular meaning of autonomy in this context requires to be specified since it might be misleading. In general autonomy is the capacity to be one's own person, to live one's life according to reasons and motives taken as one's own and not the product of external forces [3]. Moreover, autonomy is general and difficult to univocally define in order to apply it to information machines.
- We propose to intend the concept of autonomy as a *metaphor*, when used in connection with machines. Even in the case of software agents - where it is an important building block of their definition [5] - autonomy is intended in a narrower sense than in the case of humans. For software agents, autonomy

is the ability to work (rather) independently from the human user and cannot be defined absolutely, but only relatively to the goals and tasks required to the machine. One may object that it is often the case when the results of the operations performed by machines are not known in advance by human designers and users. However, this is radically different from saying that machines are autonomous and independent subjects that decide what to do. Therefore, machines should be regarded only as partially autonomous tools.

- We remark in general that the terms that are strongly humanly connoted (such as autonomy) should be adopted only metaphorically when referred to machines. The same holds for the term 'ethics' in connection to the noun 'machine'. In machine ethics [2], ethics should be adopted in a metaphorical way. It would be too narrow and difficult to make ethics a list of desiderata to be implemented in a machine. Ethics is a general category without a univocal definition. Therefore it could reveal very difficult, if not impossible, to make ethics precise enough to be programmed in order to create machines capable of ethical deliberation as humans are.
- For this reason we suggest, rather than trying to add an ethical dimension to some classes of machines, to extend the traditional ethical framework. This is not to say that we underestimate the problems involved with the new roles covered today by information machines. Instead, we are proposing to cast these problems within human ethics (enlarging existing categories) and not to introduce ethical behaviors in machines. This approach will impact on current ethical theories, promoting the discovery of new problems and the reshape of the theories themselves.
- In conclusion, we are aware of the importance of addressing the wide range of ethical problems that arise from the increasingly complex use of information machines. We think that this should be done with a very clear perspective in mind: the centrality of human beings and the differences between humans and machines. This point can be made clearer with an example. Currently, two human beings can communicate in a totally computer-mediated way. This kind of communication is very far from traditional *vis-à-vis* communication: it can involve new norms and ways of behavior, it can promote novel communication ways, and it can require sophisticated tools to overcome geographical barriers. However, it is always a communication carried on by human beings adopting technological instruments for their purposes, a process where humans remain the starting and the ending points.

## References


[1] Amigoni, F.; Schiaffonati, V.; and Somalvico, M. 2002. Multiagent System Based Scientific Discovery within Information Society. *Mind & Society* 5(3):111-127.

[2] Bynum, T. W. 1985. *Computers and Ethics*. Oxford, UK: Oxford University Press.

[3] Christman, J. 2003. Autonomy in Moral and Political Philosophy. In *The Stanford Encyclopedia of Philosophy (Fall 2003 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2003/entries/autonomy-moral/>.

[4] Gillies, D. 1996. *Artificial Intelligence and Scientific Method*. Oxford, UK: Oxford University Press.

[5] Wooldridge, M., and Jennings, N. 1995. Intelligent Agents: Theory and Practice. *Knowledge Engineering Review* 10(2).