Moral Machines:
ICTs as Mediators of Human Agency

Jos de Mul
Erasmus University

Abstract
In spite of the popularity of computer ethics, ICTs appear to undermine our moral autonomy in several ways. This article focuses on the 'delegation' of our moral agency to machines. Three stages of delegation are distinguished: implementation of moral values and norms in the design of artefacts, delegation of moral means to machines, and delegation of both moral means and goals to machines. Second, it is argued that the 'outsourcing' of moral agency does not necessarily lead to the undermining of our moral autonomy, but might enhance it as well.

Keywords: ICTs, moral autonomy, man-machine interaction, moral mediation

> The 'success' of ethics, ranging from her prominent role in various ethical commissions to her generalized status as an adjective (even banking has an ethical dimension), leads me to believe that the problem isn't so much that we *lack* ethics. What if this call for ethics itself would be part of the problem? What if the drawback of our deficit of ethics is precisely a *surplus* of ethics? What if the kind of ethics we desire helps mask the problems we are confronted with, rather than aid in facing them? -Rudi Visker (2003)

In the last few decades various information and communication technologies (ICTs) have rapidly been on the increase in almost all societal domains and in the world of our everyday lives. Human communication, financial interchanges, the production and distribution of goods, services, healthcare, military operations, science and culture have all become irrevocably intertwined with worldwide, ever more mobile computer networks. The fact that by now a significant portion of our human agency is supported, mediated or even replaced by computers and computer networks, also calls forth several moral questions.

On the face of it, it may seem as though we are dealing with often very old issues in a new guise. Of all the forms of moral behavior that we deem praiseworthy and reprehensible – assisting one's fellow human beings by word and deed, relieving distress, offering comfort, protecting the weak, cheating, stealing, and sexually abusing, intentionally harming or killing other persons – computer mediated versions have come into being over the last decades, which do not seem to be distinguishable morally from their 'real life' counterparts. In terms of moral appraisal it seems irrelevant whether an employee steals from the physical cash register of his employer or funnels off a fortune to a foreign bank account by virtual means. And our praise for the person who devotes himself to the fate of political prisoners will not depend on whether this person collects autographs in the street or attempts to gain support through an email campaign. Similarly, it seems morally irrelevant whether (child) pornography is created or disseminated by analogue or digital means. Also, many social issues relating to information technology do not seem to diverge in fundamental ways from more classical issues regarding distributive justice: think for instance of the 'digital divide' that looms between citizens or countries with and without access to new technologies.

However, on closer inspection we discover that the use of new technological means is never neutral. For one, computer mediated actions generally have a larger degree of anonymity and 'moral distance' than everyday actions, which increases the temptation to behave improperly. Moreover, in many cases the moral effects of an action are massively increased in scale. While the consequences of a classical act of vandalism are generally limited to a single bus shelter or a particular phone booth, the maker of a computer virus can damage millions of computers worldwide within just a few hours.

ICTs also make available new dimensions of acting, which call forth ethical questions that were not raised before. Is sexually harassing another person in a computer mediated environment – a chat room, an online game, a virtual world such as Second Life – of the same order as sexual harassment in 'real life'? And if we find that harassment is also morally reprehensible in virtual environments and want to punish it, should we only punish the virtual person (for instance by limiting the freedom of movement of the avatar – the digital representation of the person –, limiting access to the virtual world, or even denying access entirely), or should we also punish the real person in the actual world (by giving him a fine, or putting him in jail)?[1] Is creating child pornography morally reprehensible in the same way when the 'images' are photo-realistic and indistinguishable from the real, yet made entirely with the help of computers, instead of by actually abusing children? Are fake images such as these more closely related to traditional pornographic representations, or rather to the literary descriptions of such abusive acts, as we know them from instance from Nabokov's novel *Lolita*?

Whereas the issues in the last examples are still relatively close to traditional moral questions, in the case of the design and application of artificial intelligence, for instance in the use of expert systems or the linking between such systems and the human body, oftentimes moral questions arise that seem to have no historical counterpart. Is it morally permissible for an expert system to conduct a medical diagnosis or to pass legal judgment? What are the ethical merits of an expert system that doesn't allow the human user to perform morally inadmissible actions? Can we also hold a self-learning expert system morally accountable or responsible for its judgments? Will we ever have to compose a universal declaration of the rights of artificial intelligences? Should we implant information technologies into the human body? Hardly anyone will have moral objections regarding a pacemaker, but what of a chip that regulates the hormone production in the brain to reduce aggressive behavior or create a constant sense of happiness?

While these last questions are currently only of interest to science fiction lovers and those who enjoy philosophical thought experiments, many of the themes mentioned before – informational privacy, virtual child pornography, digital divides, the merits of medical and legal expert systems – have found their way into public debates. In addition to medical professionals, lawyers, politicians, hackers, activists, social scientists, and journalists, ethicists, too, have wholeheartedly joined in these discussions. Computer ethics in the process has rapidly grown into a popular and widely practiced ethical sub-discipline with its own classics, handbooks, journals, conferences and advisory boards (Bynum 1998).

The interest of these ethicists is not surprising. After all, in contrast with classical IT, which mainly aims at controlling non-human nature, ICTs are means for human communication and interaction. In that function it touches upon almost all aspects of human agency. Especially in those cases in which moral questions arise that are incomparable to traditional moral problems, a need emerges for codes of conduct that tell us what we should do or what we should avoid with respect to the interests and rights that we ourselves and others have. However, approaching the aforementioned questions from a moral perspective is not free of problems. After all, ICTs do not

merely call forth old and new ethical questions, but they also create a number of difficult obstacles when we attempt to pose and answer these questions using traditional ethical concepts.

Relativizing ethics

Several aspects can be distinguished in the problematization of traditional ethics that is caused by information technology. First of all, ICTs contribute significantly to the fragmentation and pluralization of society, which confronts the average citizen with various, and often conflicting, norms and values. Of course information technology is not the only, and perhaps not even the most important cause of this new moral obscurity. The globalization of the economy and the migration patterns that have accompanied this globalization, and the multicultural influences that migration and globalization have had on societies, have similarly contributed to the fact that more and more people in our society are confronted with diverse patterns of norms and values. However, one could claim that ICTs play a crucial role in this process, because these technologies, which have partially also incorporated the (function of) classical media such as newspapers and televisions, are media *par excellence* through which many of the aforementioned socio-economic, political and cultural developments not only occur, but in which they are also represented, communicated and discussed. In these discussions ethics can no longer appeal to a self-evident normative framework that is shared by everyone. The call to restore 'the' traditional values and norms shows a desire to return to a mono-cultural society that is as nostalgic as it is unrealistic. Of course, one may declare the project of the multicultural society bankrupt – as have done, for example, in my home country the Netherlands – and advocate a more or less enforced adjustment on the part of foreigners and immigrants to 'the' Dutch values and norms. But these values and norms appear to have lost their universal validity and unifying force to a considerable degree even for the 'original inhabitants'. Moreover, in these debates it is rarely made explicit which (and whose) values and norms are under discussion exactly. A factual pluralism of values and norms does not automatically relativize ethics per se, but it does weaken her position, which was not too strong to begin with, in the heterogeneous social debates, which are all too often dominated by economic and political interests and one-dimensional emotions fueled by the media.

A second obstacle that confronts ethics in light of information technology is the fact that the new questions that arise in her wake not only show a *vacuum of action* and a need for new codes of conduct, but also regularly give testimony to a *conceptual vacuum* which frustrates the formulation of new codes of conduct (Moor 1985). He who tackles the question of whether automatically archiving all e-mail transactions in an organization does or does not affect the privacy of employees, quickly realizes that a notion such as 'privacy' does not have the same meaning anymore in the age of information technology compared to the one it had in the pre-informational era. And when actions epistemically depend on a computer system, for instance because the (social) reality in which one acts only exists as a computer representation, this entails that a fundamental notion such as 'responsibility' gains a different meaning. Similarly, when one avatar sexually harasses another in a virtual environment and we decide to punish, the first and foremost question that arises is what sexual harassment and punishment consist of exactly in virtual worlds. In all of these cases we have no casuistry that can build on unproblematic normative postulates and principles, precisely because information technology fundamentally brings our moral postulates and principles up for discussion. This may be quite a challenge for the ethicists (I suspect that this is one of the causes of the significant interest in computer ethics), but it also entails that, more often than not, ethicists who participate in public debates generate more questions than they answer. ICTs create a moral vagueness and ambiguity that, and while this generally doesn't stop us from acting, nevertheless frustrates the moral validation of these

actions. And this moral vagueness and ambiguity may generate an ontological and normative pollution that can turn out quite disturbingly.

While in the previous examples information technology touches upon the fundamental concepts that a subject of moral judgments and actions uses (or would like to use), when software agents and expert systems are applied to enforce moral behavior (or to block unwanted behavior) or to take moral decisions, then this means that information technology also undermines the moral stance in a third, very fundamental way. In those cases we seem to be not only epistemically dependent on an expert system, but also morally. This would imply a fundamental deflation of our moral autonomy, that is, our ability to *freely* set our moral code of conduct *for ourselves*. In the next paragraphs I will look more closely at this third way in which information technology relativizes the ethical stance.

Before I do so, I want to make a short remark on the limits of the moral autonomy of human beings. Moral autonomy is always more of an (inviting and regulative) ideal than a reality. It is always factually and normatively finite. Its factual finitude relates to the fact that actual limitations have been set on the (modern, liberal) notion of autonomy regarding the (levels of) independence, freedom and rationality that were presupposed in that notion. We always depend on the culture and the age in which we grow up and become socialized, our freedom of choice is finite and relative, and we are rational only up to a certain extent. The normative finitude or our moral autonomy relates to the fact that our rights for non-intervention are limited, in part because of the principle of detriment that states that the government may, and must, intervene when strategically acting individuals harm the rights of others. The question that I will attempt to answer is not *whether* factual or normative limits have been placed on our moral autonomy, but rather *in which ways* information technology has an impact on these limits. I will address the factual question of whether information technology limits our moral autonomy, if so, in which ways, and I will address the normative question of how to value these matters.

Acting machines

Because of its programmability the computer is often called a universal machine. Whereas the classical machine is a representation of one specific program, the computer is 'a mechanism that realizes the physical representation of every installed program as one of her possible operating procedures' (Coolen 1992, 39). Because of this fundamental versatility and flexibility the computer can be applied in innumerable ways, also in the realm of moral action. It is impossible to give even an overview of these possibilities. This is why I will limit myself to discussing what I view as the three fundamental types of delegation of moral action and judgment to computers. I present these three fundamental types in an order that displays the gradual decrease of moral autonomy for man and the steady increase of moral autonomy for computers.

The *first type* of 'delegation' of morality to the computer revolves around the implementation of moral values and norms in the design of the computer program by human beings. In practice most of the time this involves a process of negotiation – "a semiotic power struggle" (Bijker 1995) – between, among others, clients, financiers, designers, governments and (potential) users. By prohibiting morally reprehensible actions and/or enforcing morally desirable ones, the computer system invites the user to display morally desirable behaviors. An example of such programs that of *privacy enhancing technologies* (PETs), which are embedded in information systems. These programs aim at protecting the privacy of individuals by eliminating or reducing personal data or by prohibiting the processing of such data whenever processing is unnecessary and/or unwanted, without interfering with the functionality of the system (Borking and Raab 2001).

This kind of moral delegation to the computer in principle cannot be distinguished from other forms of the moralization through objects, such as the placement of tourniquets at the entrance to the subway. In the latter case, too, an ethical value or norm is implanted by man into the technical design, in order to enforce moral (and legal) behavior. Strictly speaking, in both cases we cannot speak of moral actions or decisions by a non-human, intentional and (self)conscious, moral actor. Obviously, we find instances of the *technological mediation* of morality in these examples, but not of a morally acting technological agent. Both the goal (the values to be realized) and the means (the norms required for their realization) are implanted into the technology by man.

Viewed from a utilitarian perspective, such a 'moralization of things' is often superior to other strategies that focus on the intentionality of users, such as moral education, information campaigns and legislation with accompanying sanctions. While extensive information campaigns did not notably contribute to a decrease of fare dodging in the Amsterdam subway, the placement of a simple gate– that could be passed also without a subway ticket – in the nineties led to a reduction in fare dodging from 35% to 15% (Achterhuis 1998, 369). The moralizing tendencies of information technologies raise various questions and objections. Here, too, we are confronted with conceptual confusion and ambiguity. Can one argue that the subway traveler has become more moral because of the gates? Is it ethically relevant whether the desired situation is brought about by upbringing, training, disciplining practices, a ticket inspector, an object (gate), a text (legal code), or a computer program (PET)? One of the fundamental objections raised is the idea that the delegation of morality, as discussed above, would have a negative impact on the moral autonomy of human beings. When individuals *cannot* make moral mistakes, they cannot acquire moral wisdom, nor grow to become morally refined human beings. Moreover, technological design may have unforeseen and unwanted moral effects – effects that appear to undermine human autonomy as well. In the first section I mentioned the consequences of the increase in anonymity and 'moral distance' created by information technology. Sometimes technologies may even unintentionally induce behaviors that are the reverse of the values and norms that are embedded into the technological design. The answering machine was designed to increase individuals' availability, but in fact in practice it is often used to be selectively unavailable. An unintended effect of mobile phones and security cameras is, for instance, that individuals witnessing a violent crime are less inclined to intervene. Instead, they are more predisposed to merely call the police's emergency number, or even to not do anything at all, because they assume that the police will be notified by the cameras. Practical issues that arise relate to questions such as: 'who is responsible for the design of moralizing artifacts and software?', 'on whose authority are such artifacts and systems designed?', and 'to which types of democratic control are they submitted?'.

Whereas in the first type of moral 'outsourcing' both the goal and the means are implemented by human beings into the technical design, and the information technology appears to be responsible only for the unintended effects, in the *second type* of delegating morality finding the adequate means is left to the program itself. In many current-day cars the brake system contains various chips that decide which wheels will be slowed down and how powerfully this should be done in each individual braking maneuver; sometimes they may even correct the position of the wheels whenever that is deemed necessary. The driver decides what must happen (the goal), but how that goal is to be accomplished (the choice of the means), is left to the computer program – within certain limits defined in the design. In this case, the issue relates primarily to technical means. In the case of legal and medical expert systems there is also a delegation of finding normative means. The medical system APACHE III is used to predict the chances that patients in the intensive care unit of a hospital may die as a result of the diseases they suffer from.[2] It turns out

that APACHE III's predictions are significantly more accurate than those made by doctors. The reason why this is so, is because doctors are (often unconsciously) inclined to factor in the patient's age in a disproportionate fashion. In all likelihood, this is done both on medical and on normative grounds – younger patients are viewed as more resilient and when they survive they stand a better chance to have a long life than elderly patients. For this reason the survival chances of younger patients are often overestimated, and younger patients frequently get precedence over others when there is a shortage of beds in the intensive care unit. APACHE III appears to be much better at predicting, partially because the system, in analyzing the submitted statistical data, ascribes much less weight to the factor of a patient's age than do human doctors.

In APACHE III's case we are also dealing with a delegation of morality – this time not by enforcing or prohibiting specific behaviors but by letting the machine rationally deliberate the means that precede action – a rational deliberation that, from a utilitarian perspective, is superior to that of mere humans. After all, when young patients, with a considerable likelihood of dying, get precedence over older patients with a higher life expectancy (but who also run more of a risk to die outside the intensive care unit), then this means that the goal, i.e., saving as many years of life as possible, is realized only in suboptimal form.[3]

It seems obvious that the moral autonomy of human beings is limited to a more radical degree in this second type of delegation as was the case in the first type. The doctor, indeed, may in principle ignore APACHE III's prediction and carry out his own plan, but we may wonder whether he would actually do so. After all, in the current claim culture he runs the risk of liability if he follows false intuitions. Moreover, the question arises of whether he ought to be held morally responsible, because he chose to ignore the predictions made by the system. In contrast, the moral autonomy of APACHE III appears to expand because of this reason. Its expansion in this case is not so much metaphorical, in the sense that the technical actions of man may at times have unexpected and unforeseen side-effects, but rather literal, in the sense that the information technological system itself gains particular characteristics that we ascribe to moral agents. While the system is not (self)conscious and has obtained its values from an external source (in which it resembles a child that receives moral training, by the way), it does develop something bordering on an unconscious, immanent intentionality with respect to the means.[4]

The problem of the corrosion of human autonomy is sharpened when we pose the question whether human beings ought to always be allowed to override or ignore a system's decisions. Some airplanes, such as the Boeing 777, fly entirely 'by wire', that means on autopilot. The pilot can resume control whenever that is required for some reason. This seems to be a sensible solution, considering the fact that the 'real existing technology' is far from infallible. There are numerous examples of technical and moral disasters that were caused by the *mal*functioning of computer programs (such as the string of accidents of the Airbus A320 that were caused by bugs in the autopilot), or disasters that were caused – tragic irony! – because the program *did*, in fact, function properly (as was the case, for instance, in the collapse of the stock market in 1987, when the perfect functioning of the automated sales system for stocks caused a dramatic drop in the stock exchange). But on the other hand, since man is far from infallible as well, it might be sensible to allow the program to override the human agent in turn – to correct when human beings inadvertently (panic) or purposefully (terrorist attack) execute actions that will almost certainly lead to disaster.

The *third* type of moral delegation takes things one step further yet and arises in relation to self-learning expert systems, which are based on genetic algorithms and neural networks, and which don't merely decide on what means to use – as does APACHE III – but also set the moral goals,

the moral values that ought to be achieved. An intriguing – partially realistic, partially imaginary – example of such a system is provided by Eric Steinhart in an article on emergent values in automatons (Steinhart 1999). He describes a *power grid*, an intelligent network for electricity consisting of power plants, switches, transformers, electronic devices, computers, and human manufacturers and consumers. The wires that connect these nodes do not merely transport electricity but also information. The system is designed to distribute electricity as reliably and efficiently as possible. Reliability and efficiency are values that are embedded into the system by human beings. In the first developmental phase, the power grid does not distinguish itself from the automated brake systems and expert systems described above. While human beings prescribe *what* the network ought to accomplish, it is the network itself that largely decides *how* it will go about realizing these values. In Steinhart's network *software agents* negotiate with one another to purchase electrical energy as cheaply as possible for their owners. Say you would like to toast some bread in the morning, but the power prices are very high at that moment, then the toaster's software agent will attempt, for instance, to borrow a few watts from the washer or the heating system. If it succeeds, you can make the toast; if it doesn't, you'll have to pass up on your toast this morning, but at least you will have accomplished the prescribed goal (in this case: cutting back on your energy bill or saving the environment).

We can also conceive of designing this kind of system in such a way that it is self-organizing and will start to formulate its own goals. For instance, we could imagine that it would develop new patterns of distribution that safeguard the allotment of power in the longer term (an emergent value) at the expense of the instant satisfaction of needs of current-day customers. In this case the delegation of morality is taken one step further again, and the moral autonomy of the system is increased once more. The system now not only limits itself to a choice regarding the (moral) means, but also decides on the (moral) goals. Such an intentional system still lacks (self)consciousness, but despite this fact it has become an agent with a considerable degree of autonomy:

> The power grid seems to be smart or intelligent, but it's not: it's just a complex self-organizing system. It's more like an immune system than like a mind. It isn't conscious or self-conscious. But thought is not necessary for action. The power grid does not think, but it acts. It acts on its own. It's action is involuntary and mindless. But it is nevertheless political, since it is free and it affects us (Steinhart 1999, 158).

One may justifiably wonder whether we could call the network that Steinhart describes a moral agent that is responsible, or may be held accountable for its actions. The answer to this question depends, in part, on the way one answers another question, that is, whether (self)consciousness is a necessary requirement for an entity's ability to act morally. If we answer the latter question affirmatively, then much depends on whether or not self-organizing systems will ever develop consciousness. While to my mind this cannot be ruled out, this issue is not really relevant for the question we are addressing here. Even if we are unwilling to call the network a moral agent in the human sense, nevertheless we have to admit that it is an autonomously acting system that has a more or less imperative influence on the moral actions of its human users.

ICT as moral mediator

On the face of it, the progressive 'outsourcing' of morality appears to confirm the relativization of ethics that I discussed in the second paragraph. While the human moral agent remains responsible for the development and implementation of moral rules in the case of the first type of delegation, in the second type he does so only in a derived sense, and in the third, most extreme type of delegation the human moral agent seems to loose involvement in the morality of the information system anymore completely.

There are, however, valid reasons to doubt whether the delegation of morality to computer systems even of the third, radical kind, truly undermines the moral agency of human beings. That conclusion, after all, can only be drawn if one, first of all, views computers and human beings as two strictly separate entities, and on top of that, if one views formulating moral goals as a matter that should be exclusively reserved for human beings, and technologies merely as the providers of neutral means. Only on the basis of those premises can the delegation of goals to a computer system be considered a radical loss of moral autonomy. However, both of these premises are problematic indeed.

Following the philosophical anthropologist Plessner we could claim that man is "artificial by nature". Man is characterized by a fundamental neediness, which precedes and underpins every subjective need, craving, drive or impulse. In this interpretation man only becomes himself through the supplemental means of technology and culture. These supplements are not merely instruments for humans' survival, but rather an "ontic necessity" for man (Plessner 1975, 396). This means that despite their alterity these supplements are part of the self. Human life is enacted in the heterogeneous ensemble of the organic basis and its supplements. Thus, while an 'alpha-technology' such as writing is not really part of the organism, it is nevertheless indissolubly connected with the cognitive structures of lettered mankind. The delegation of memory to supplementary writing does not imply that writing is exclusively 'outside' of human beings. On the contrary, it is interwoven with body and mind in a multitude of relations: it directs our hand, attitude, memory and imagination, it opens a whirlwind of new worlds, it widens our cognitive repertoire and our range of possible actions, and in the process it enlarges our freedom. In contrast, that which we call 'human' is not entirely 'outside' of writing. Writing is an integral part of what constitutes human being(s). And as a matter of fact, this applies to any technical or cultural artifact. Man has been a cyborg ever since the moment he started using tools, partially an organism, partially technology and culture (cf. De Mul 2003).

This is an idea that has met with resistance in western philosophy from the outset. In *Phaedrus* Plato critically discusses the delegation of human memory to writing, in a dialogue between the Egyptian king Thamus and Socrates. Writing, says Thamus,

> will create forgetfulness in the learners' souls, because they will not use their memories; they will trust to the external written characters and not remember of themselves. […] You give your disciples not truth, but only the semblance of truth; they will be hearers of many things and will have learned nothing; they will appear to be omniscient and will generally know nothing; they will be tiresome company, having the show of wisdom without the reality (Plato 2007, 275a).

Plato takes a stance that we often find in discussions on the delegation of cognitive tasks to computers. In those discussions, too, it is argued that the delegation – in this case not of the products of thinking, but of the rational and moral thought processes themselves – to a separate

machine undermines human autonomy. In both cases the fact that the technical 'means' forms an integral part of our distributed cognitive structure is ignored, but, more importantly, the fact that in this structure new, 'typically human' capabilities are created is overlooked.[5] The delegation of memory to writing relieves our cognitive structure of the task of remembering and opens up opportunities for the development of new modes of rationality and morality. We may expect similar events to occur in the case of delegating our rational and moral capabilities to computers and expert systems. This delegation does not entail an impoverishment of human autonomy, but rather an expansion thereof. In all cases, however, human autonomy is *supplemented autonomy*, characterized by an insolvable alteriority and exteriority.

In Bruno Latour's work we find a similar perspective regarding the moral implications of technological mediation. Latour rejects the assumption that human ethics formulates moral goals, whereas technology merely supplies the means to realize these goals. Technology always provides us with a *detour* towards the goals we want to reach:

> If we fail to recognize how much the use of a technique, however simple, has displaced, translated, modified, or inflected the initial intention, it is simply because we have *changed the end in changing the means*, and because, through a slipping of the will, we have begun to wish something quite else from what we at first desired. […] Without technological detours, the properly human cannot exist. […] Technologies bombard human beings with a ceaseless offer of previously unheard-of positions – engagements, suggestions, allowances, interdictions, habits, positions, alienations, prescriptions, calculations, memories. (Latour 2002, 252).

In his attempt to restore technology's 'ontological dignity' Latour regularly discusses the moralization of objects in his work. In the article quoted above, he admits that he may have "exaggerated somewhat" when he talked of the "tragic dilemmas of a safety belt" (Latour 2002, 254). For as far as human values and norms are implanted into technologies such as the safety belt or privacy enhancing technologies, this relativization is correct. I noted above that in these cases we wouldn't speak of autonomous moral agents (unless one uses the term as synonymous to that which is unforeseen and/or unintentional, from the human perspective). However, in the examples of APACHE III and the power grid we saw that there is, in fact, some level of autonomy and intentionality in those cases. But even here the autonomous structures are not entirely separate from ourselves; they can become part of our extending cognitive structure. If that is so, then we incorporate them into the heterogeneous field of forces, to which our emotions, our rational considerations, our urges, disciplines and our technological and cultural artifacts also belong.

The delegation of our memory to writing and of certain thought processes to the computer has led us, and enabled us, to develop new cognitive capabilities. These, in turn, have made it possible for us to incorporate this external memory and artificial brain into our own cognitive structure. Mindful of Plato's critique, the task we need to set for thinking is to preclude that the delegation of our memory will result in strange, dead letters.

In a similar fashion, the relief of our 'moral organ' by the delegation of moral capacities to machines and systems challenges us to develop new moral capabilities that will enable us to incorporate the processes that take place in those machines and systems into our human morality. The challenge for ethics is to prevent the heterogeneous field of forces of living morality from calcifying into a blindly functioning system. We cannot accomplish this task by attempting to regain 'full' control of all aspects of moral reasoning or desperately trying to force 'our own'

goals onto the technology once more. To do so would be a grave failure to appreciate not only the nature of technology, but also of the unprecedented complexity of our current societies. Rather, ethics ought to 'dislocate the tracks' "[w]herever we want to go fast by establishing [such] tracks so that a goal can race along them whistling like a high-speed train' and to constantly call attention 'all the crossroads and lost sidings" (Latour 2002, 257).

We cannot deny the fact that the delegation of our morality to technology is not without dangers. This process is accompanied by a fear – as if by a shadow – of a technology that turns against us – either by functioning (too) well or by malfunctioning – and that destroys our humanity or even the human species as such.[6] That risk is real and may never be forgotten. However, we also shouldn't forget that it was this same technology that gave us – among other things – our humanity and our morality several million years ago. This fundamental ambiguity of the technological supplement inevitably reflects back on our humanity: 'Inhumanity is not bound up with a specific age or a particular historical grandness, but is rather a possibility that is given in man: to ignore himself and his equals' (Plessner 1982).

References

Achterhuis, H., ed. 1998. *De erfenis van de utopie*. Baarn: Ambo.

Bijker, W.E. 1995. *Democratisering van de technologische cultuur.* Inaugural address. Maastricht: University of Maastricht.

Borking, J. and C. Raab. 2001. "Laws, PETs and other technologies for privacy protection." (Available from http://elj.warwick.ac.uk/jilt/01-1/borking.html.

Bynum, T.W. 1998. "Global information ethics and the information revolution." In *The digital phoenix : how computers are changing philosophy*, edited by T. W. Bynum, J. H. Moor and American Philosophical Association. Committee on Philosophy and Computers. Oxford ; Malden, MA: Blackwell Publishers.

Coolen, M. 1992. *De machine voorbij. Over het zelfbegrip van de mens in het tijdperk van de informatietechniek*. Amsterdam: Boom.

Latour, B. 2002. "Morality and Technology: The End of the Means." *Theory, Culture, Society* 19 (5/6):247-60.

MacKinnon, R.C. 1997. "Punishing the persona: Correctional strategies for the virtual offender." In *Virtual Culture: Identity and Communication in Cybersociety*, edited by S. G. Jones. London: Sage Publications.

Minsky, M. 1985. *The Society of Mind*. New York: Simon and Schuster.

Moor, J.H. 1985. "What is computer ethics?" In *Computers and Ethics. Special Issue of Metaphilosophy*, edited by T. W. Bynum.

Mul, J. de. 2003. "Digitally mediated (dis)embodiment. Plessner's concept of excentric positionality explained for Cyborgs." *Information, Communication & Society* 6 (2):247-266.

Okrent, M. 1996. *Why The Mind Isn't a Program (But Some Digital Computer Might Have a Mind)* Available from http://ejap.louisiana.edu/EJAP/1996.spring/okrent.1996. spring.html.

Plato, P. 2007. *Phaedrus*. Fairfield, IA: 1st World Library/Literary Society.

Plessner, H. 1975. *Die Stufen des Organischen und der Mensch. Einleitung in die philosophische Anthropologie*. Vol. IV, *Gesammelte Schriften*. Frankfurt: Suhrkamp.

Plessner, H. 1982. "Unmenschlichkeit." In *Mit anderen Augen: Aspekte einer philosophischen Anthropologie*. Stuttgart: Reclam.

Steinhart, E. 1999. "Emergent values for automatons: Ethical problems of life in the generalized Internet." *Ethics and Information Technology* 1 (2):155-160.

Visker, R. 2003. Nooit meer slapen. Over ethiek en onverantwoordelijkheid. In *De verleiding van de ethiek. Over de plaats van morele argumenten in de huidige maatschappij*, edited by I. Devisch and G. Verschraegen. Amsterdam: Boom, p. 22-42.

Endnotes

1 In an interesting essay on the punishment of virtual vices MacKinnon uses a number of cases to show that the punishment in similar cases is often different (MacKinnon 1997).

2 See the Apache Medical Systems Inc. website: http://www.openclinical.org/aisp_apache.html.

3 Apache calculates -- in a properly utilistic fashion the number of years of life, rather than 'lives proper'. This means that from Apache's perspective it is better to safe a 60-year old person when he or she gains another ten years of life, rather than a 20-year old who will only gain another 5 years of life through intensive care. These calculations, of course, are based on the past average of expected extra years of life: damage of type X means that

the expected number of years of life is 10 years for a 60-year old; damage of type Y in a 20-year old entails 5 years of life etcetera.

4 Compare this to Mark Okrent's argument (which was inspired by Heidegger's critique of the Cartesian emphasis on the role of consciousness) that we may ascribe intentionality to entities even when we cannot necessarily ascribe (self)consciousness to them (Okrent 1996).

5 This distributed character does not merely apply to the distribution of the cognitive structure between human beings and their technological and cultural artifacts, but also applies to the naked human mind itself, which is not a unity, but consists of many heterogeneous sub-systems (Minsky 1985).

6 Kubrick has majestically portrayed this fear in his film *2001: A Space Odyssee*, in which the self-learning on-board computer HAL turns against the crew of a spaceship so that the accomplishment of its mission, reaching the planet Jupiter, is not endangered.