Information, Ethics, and Computers: The Problem of Autonomous Moral Agents

BERND CARSTEN STAHL
De Montfort University, Centre for Computing and Social Responsibility, Leicester LE1 9BH, UK;
E-mail: bstahl@dmu.ac.uk

Abstract. In modern technical societies computers interact with human beings in ways that can affect moral rights and obligations. This has given rise to the question whether computers can act as autonomous moral agents. The answer to this question depends on many explicit and implicit definitions that touch on different philosophical areas such as anthropology and metaphysics. The approach chosen in this paper centres on the concept of information. Information is a multi-facetted notion which is hard to define comprehensively. However, the frequently used definition of information as data endowed with meaning can promote our understanding. It is argued that information in this sense is a necessary condition of cognitivist ethics. This is the basis for analysing computers and information processors regarding their status as possible moral agents. Computers have several characteristics that are desirable for moral agents. However, computers in their current form are unable to capture the meaning of information and therefore fail to reflect morality in anything but a most basic sense of the term. This shortcoming is discussed using the example of the Moral Turing Test. The paper ends with a consideration of which conditions computers would have to fulfil in order to be able to use information in such a way as to render them capable of acting morally and reflecting ethically.

Key words: autonomous moral agent, ethics, meaning, information, morality, responsibility

1. Introduction

Ethical theory and moral practice originally refer to human behaviour. In order to act morally and to reflect on this behaviour from an ethical point of view one needs information. This information concerns the factual state of affairs and also the normative evaluation of these facts. Information is thus a necessary condition of ethics and morality. At the same time we call computers information processors. Given the link between ethics and information and the information processing capabilities of computers one can ask what role computers can play in ethics. Are computer able to use information to act morally or reflect ethically?

In this paper I will attempt to give an answer to these questions. I will ask whether computers can use information in order to act morally in a way similar to humans. In the next step I will look at the concept of information. The definition and description of information will aim at finding out in what respects information has an impact on ethics, particularly cognitivist ethics. The emphasis of this discussion will be on the question whether computers can be Autonomous Moral Agents (AMAs). One way that has been suggested as a means of determining this

Minds and Machines **14:** 67–83, 2004.

is the Moral Turing Test. I will ask what the limits of such a test are and whether it is possible for computers to pass it. The result of this analysis will be that computers in their current form do not appear to be candidates for moral agency, mainly because they do not capture the meaning of the data they process. The conclusion will ask whether this result is a fundamental one and which changes in computers would be necessary in order to render them autonomous moral agents.

2. Information and Computers as Autonomous Moral Agents

In everyday life in most industrialised countries it happens quite frequently that human agents interact with machines. While this has happened in a more trivial way ever since humans first invented tools, the quality of these interaction changes due to the increasing use of computers. The average citizen will today frequently be forced to interact and communicate with computers in ways we used to interact and communicate with humans. From the automated teller machine to a synthesised voice in telephone directory inquiry, machines act autonomously, meaning they interact with humans without specific instructions for particular interactions. This type of interaction can be of a moral nature. A computer can produce statements about a human being or initiate actions that affect the moral rights and obligations of that human being. Machine and computer actions can therefore have a moral quality. The question now is whether computers can also be seen as moral agents, whether they can be ascribed moral responsibility.

2.1. MACHINES AS SUBJECTS OF MORAL RESPONSIBILITY

Traditionally, most moral philosophers would deny the possibility of machines being subjects of responsibility (Jordan, 1963; Lenk, 1994, p. 84). “Moral responsibility is attributed to moral agents, and in Western philosophy that has exclusively meant human beings.” (Johnson, 2001, p. 188). There are many possible explanations why philosophers believe that only humans or persons can be morally responsible. One of them is a metaphysical assumption that only humans have the status of being that allows responsibility ascription (Stahl, 2000). This in turn can be justified by the fact that in order for responsibility ascriptions to make sense, the subject must fulfil a number of conditions such as cognitive and emotional abilities, some knowledge of the results of actions as well as the power to change events. Humans can be guided by fear of sanctions or hope of rewards. Humans can recognise others as equal and humans are social beings who rely on morality. Another argument against a moral status of machines is that it can be used as an excuse for people to evade their responsibility (Grint and Woolgar, 1997).

On the other hand, there are good arguments for considering the possibility of ascribing an independent moral status to computers. As indicated earlier, computers often play roles in social interaction which have a moral nature. Computers can be made part of explicit moral decision making (Gotterbarn, 2002). Furthermore, the aim of computer development is often the creation of autonomous agents. For many technical purposes computers or robots need an increasing amount of autonomy. The Pathfinder robot of the Mars mission, for example, had to be highly autonomous because it was simply impossible to control it real-time. Another example are software bots which act in their environment and can display moral characteristics (Mowbray, 2002). If machines can act autonomously and be involved in moral situations then the question becomes prevalent whether they can react adequately to normative problems. Further arguments for an autonomous moral status of computers and machines are that humans may not be capable (any more) to live up to ethical expectations and that only computers are able to do so (Bechtel, 1985) or the practical argument that we are simply not able to reduce machine actions to human actions and therefore need to ascribe moral responsibility to computers as a pragmatic and manageable solution (Stahl, 2001). Finally, some authors go so far as to envisage a future in which computers will be called upon to fulfil the most responsible of tasks such as that of a judge (cf. Stewart, 1997). For these and other reasons one can increasingly find philosophical arguments of the following kind: "AAs [Autonomous Agents, BCS] are legitimate sources of im/moral actions, hence A [the class of all moral agents, BCS] should be extended so as to include AAs, that their ethical discourse should include the analysis of their morality [. . .]." (Floridi and Sanders, 2001, p. 3)

2.2. AUTONOMOUS MORAL AGENTS AND THE MORAL TURING TEST

The arguments collected in the last section require that the possibility of computers as moral agents be taken seriously. This is not a fundamentally new insight and has been discussed, for example, by Bechtel (1985). The developments outlined make this question more urgent today because computers become ubiquitous in many areas. Some scholars have attempted to develop the discussion by introducing the concept of the computer as autonomous moral agent (AMA) (cf. Allen et al., 2000; Allen, 2002).

The question whether computers can be genuine moral subjects depends in many instances on the metaphysical convictions of the participants. It will therefore not be answered conclusively in the near future. One way of furthering the debate, however, would be to do empirical research that could demonstrate whether computers can display moral behaviour similar to humans. As a theoretical basis of such research Allen et al. (2000) introduce the Moral Turing Test. The original Turing test, suggested by Alan Turing (1950), aimed at the question whether computers can think. Turing recognised that this question is almost impossible to answer because the definitions of thinking vary too widely. He therefore considered the question "too meaningless to deserve discussion" (Turing, 1950, p. 442). Instead, he proposed an "imitation game". The point of the game is that an outside observer conducts a written conversation with two parties, one of whom is a computer, one a human being. The observer is charged with finding out who the computer is in a finite amount of time. If the observer fails to do so with a high enough percentage then the computer can imitate a human being and this, according to Turing, is a more interesting aspect than the question whether it can think.

The Moral Turing Test chooses a similar approach. Since it is sufficiently difficult to get philosophers to agree on what constitutes ethics and morality the chances of finding an answer to the question whether computers can act morally or reflect ethically seems moot. Instead, if we could test computers' behaviour and see whether they pass for moral beings in an independent observer's view then this could constitute a criterion that would allow us to say that they are autonomous moral agents.

> A Moral Turing Test (MTT) might [...] be proposed to bypass disagreements about ethical standards by restricting the standard Turing Test to conversations about morality. If human "interrogators" cannot identify the machine at above chance accuracy, then the machine is, on this criterion, a moral agent. (Allen et al., 2000, p. 254)

2.3. INFORMATION AND AUTONOMOUS MORAL AGENTS

The problems and questions related to computers acting as moral agents are manifold and can be addressed from many different viewpoints. One can look at them from the point of view of (moral) psychology, cognitive sciences, different ethical theories, just to name a few. In this paper, I will apply the theoretical lens of the philosophy of information to the problem. Philosophy of information has been defined as "the philosophical field concerned with (a) the critical investigation of the conceptual nature and basic principles of information, including its dynamics, utilisation, and sciences, and (b) the elaboration and application of information-theoretic and computational methodologies to philosophical problems." (Floridi, 2002, p. 137) For the purposes of this article part (a) of the definition is applicable. The paper will discuss the nature of information with regards to its impact on ethics. It will concentrate on a group of ethical theories for which information plays a central role, namely on cognitivist ethics. It will be argued that cognitivist ethics with their assumption that ethical statements have a truth-value are quite close to information and rely on information. Furthermore, if computers can obtain the status of autonomous moral agents then they will be most likely to be successful in the light of cognitivist ethical theories. In a further step the problems of this approach will be discussed and it will be shown that the concept of information used for and by computers differs vastly from the concept of information required by cognitivist ethics. This will be done by returning to the idea of the Moral Turing Test and by discussing which problems computers will run into when attempting to pass it.

The argument will be that cognitivist ethics requires an application of an abstract theory to a real situation. A simplistic view of information might state that all one has to do is take the abstract information of the ethical theory and fit the concrete information regarding the situation to it. Since computers are information processors they should be able to do so and thus pass the Moral Turing Test. What this sort of argument misses is the fact that information is never given in an absolute sense but always part of a greater universe of meaning and practice. In order to be able to process information a subject needs more than mathematical rules. It requires an understanding of the universe of meaning, which in turn requires several other factors, among them a physical existence and a being-in-the-world.

3. Information and Cognitivist Ethics

This section will begin with a definition of the concept of information. It will then describe some of the possible relationships between ethics and information. In a last step the chapter will focus on those ethical theories for which information seems to be of highest importance, namely cognitivist ethics.

3.1. INFORMATION AS DATA WITH MEANING

One of the fundamental problems of dealing with information is that it is quite hard to define. The term has many different meanings and is of importance to many different academic disciplines. This paper will mostly rely on sources from the discipline of information systems (IS). There are several reasons for this. First, the discipline of IS is by now quite well developed and has a recognised disciplinary background. Second, this background is used not only for theoretical discussions but stands in a constant exchange with organisational practice. Third, the use of computers and IT as users and processors of information is mostly driven and motivated by economic concerns and the question of computers as autonomous moral agents is therefore of central importance for the future development of IS.

In this context, information is usually part of a group of notions containing other terms such as data, knowledge, or judgment. A brief definition that in some form or other can be found in most textbooks on IS is that information is data with meaning. Data as the underlying term is often defined as "a set of discrete, objective facts about events" (Davenport and Prusak, 1998, p. 2). A similar definition of data would be: "pieces of related facts and their values" (French, 1990, p. 29). While data consists of the brute facts that are out there, information is what becomes of data once it has been transformed in a certain way. "The technical-functional notion of the relationship between data and information is that information is data that has been processed (or interpreted) in order to make it useful, sensible, or meaningful." (Introna, 1997, p. 77) Put differently, "When "data" acquires context-dependent meaning and relevance, it becomes information" (Ulrich, 2001, p. 56).

This definition of information as data with meaning has the advantage of being understandable and easy to remember. However, as some of the quoted authors note, it is also simplistic. It assumes the givenness of data which is problematic. Every real situation consists of a virtual infinity of aspects. Neither humans nor computers are able to process this infinity of potential data. What we call data is therefore always a subset of the possible facts and this subset must be chosen by some method. Data is therefore not the "brute reality" but it always already filtered and processed, it is the result of cognitive processes and social constructions.

Furthermore, the definition changes our problem of finding out what information is but it does not solve it. We would now have to understand what meaning is. Meaning, however, is a complicated term that is interlinked with social and cognitive processes, with the philosophy of mind, with metaphysics, and with epistemology. In the academic discipline of information systems these questions are frequently ignored because students can easily understand examples in which computers take huge amounts of, say, sales data and produce an understandable chart (information). For the purposes of the autonomous moral agent, however, the term "meaning" would have to be clarified.

> Philosophically speaking, information science and technology (IT) appear to have got it wrong from the start. Their conceptual and technical tools basically deal with the processing and transmission of signals (or messages, streams of signs) rather than with "information." (Ulrich, 2001, p. 59)

These considerations lead to doubts regarding the status of computers as information processors which will form the basis of the following argument in this paper and will be used to reject their moral status.

For the purposes of this paper the definition of information as data with meaning is useful because it will guide us through the discussion of some of the problems that computers encounter when they are to function as autonomous moral agents. Nevertheless, it may be helpful to try another avenue for the understanding information without defining it formally, which is what we will do in the next section.

3.2. THE PURPOSE AND METAPHYSICAL STATUS OF INFORMATION

In order to answer the question whether computers can use information in order to function morally it will prove useful to ask how and for what purposes humans use information. In the most general sense information is needed to cope with the environment (Wiener, 1954). While this may be true to a certain degree for all living organisms, humans use information to tailor their environment to their needs and to achieve their goals within that environment. "[...] information forms the intellectual capital from which human beings craft their lives and secure dignity" (Mason, 1986, p. 5). Information is thus the basis of human society and in conjunction with information technology seems to form the backbone of modern developments of society, known under the heading of the information society (cf. Mason et al. 1995; Castells, 2000; Zerdick et al. 2001). Within this society information can function in different ways, for example, as a symbol of power or rationality (Introna, 1997).

The different functions of information would allow its classification as a tool of social existence. However, information is more than that. It has a profound metaphysical status for society and for individuals. On one level, information creates the reality in which individuals live by opening possibilities and choices to them (Zerdick et al., 2001, p. 38). On a more fundamental level information creates the reality that it then describes. The reality in which we move only becomes reality by being turned into information that can be used by humans. This corresponds with the definition of information as data with meaning. We are surrounded by a potential infinity of data, by reality as it is, which we have to endow with meaning in order to be able to make use of it. This means that the constitution of information is a more complex process than discussed so far. The starting point is reality but reality does not provide us with data. Agents, whether human or not, must take their perceptions of reality in order to identify the data that is relevant. In a second step they then have to endow the data with further meaning necessary to act on it. The whole process becomes more complex because it is reflexive. Information constitutes meaning and this meaning allows us to perceive data in such a way that it becomes meaningful and produces more information. This refers to natural facts as well as social relationships. A stone only becomes a stone by being perceived as one and only then can it be put into use. While one may not need to know the name for something in order to use it, one needs information about it, one needs to be able to attach meaning to it. In the example of the stone, a caveman or a monkey may not know that it has a name but they possess the information that it can be used as a weapon or a tool.

While this constructionist approach may seem objectionable to some readers when applied to “natural facts” (which, of course, do not exist in this worldview) it should be relatively uncontroversial for social facts or relationships. A socially relevant artefact, say, an altar or a crown, only acquire their usefulness in conjunction with information about their meaning. Information in social settings thus has different levels of social functions and meanings. On a fundamental level information creates reality but it also allows us to individually and collectively function in this reality and use and modify it according to our purposes. Information also describes the reality it creates and to do so in a useful way it must be true (cf. Floridi, forthcoming). Critics might point out that this relationship of information and truth is tautological. I suspect that this is in fact the case, that we create the information which constitutes the reality which we use to assess the truth of information. We are thus caught up in the hermeneutic circle but for human beings that does not usually create any problems (cf. Gadamer, 1990). For computers, who are not part of the circle, this may be the root of the problem of meaning as will be discussed later.

These last few paragraphs contain several aspects that can be quite contentious from different philosophical positions. Some of the readers may not agree with the metaphysical importance attributed to information. However, it should be quite uncontroversial that information plays a multi-facetted role in the constitution of our life-worlds. Unfortunately, the use of the concept of information in some disciplines such as information systems tends to be quite narrow. Since information is used in information technology and information technology needs pieces of information coded in specific ways which can be expressed in mathematical formulas, information is frequently equated with numbers and generally assumed to be a quantifiable entity. This is enforced by the location of information systems research in the social sciences and the strong quantitative bias in mainstream social science research (Bloomfield and Coombs, 1992). This research trend finds its continuation in business trends where information is increasingly seen as a commodity, something that can be measured, bought, and sold (Stichler, 1998; Ladd, 2000). However, the quantification and commodification of information corresponds neither with its purpose of constituting meaning nor with its function regarding ethics.

3.3. INFORMATION AND (COGNITIVIST) ETHICS

As hinted at before, in this paper morality will be understood as the set of norms that guide our factual behaviour whereas ethics is seen to be the theory and reflection of morality. Information in the sense laid out above has several close links with ethics and morality. The philosophy of information as defined by Floridi (2002, p. 123) deals, among other things, with the way information is utilised. The nature of information as "human, expendable, compressible, substitutable, transportable, diffusive, and shareable" (Mason et al., 1995, p. 41) creates some unique ethical issues.

One important fact is that neither information nor information technology can be ethically neutral. It changes the way we perceive and construct our reality. "The word "inform" originally meant "to give shape to" and information is meant to shape the person who gets it, to make some difference in his outlook or insight" (Davenport and Prusak, 1998, p. 3). While information changes our reality, information technology changes the way we handle and understand information. Computers can, for example, "grease" information, change the speed and the reach with which it is exchanged (Moor, 2000). The different ways in which information and information technology affect our moral norms and their ethical reflection are discussed in the fields of computer ethics, information ethics, Internet ethics etc. Issues range from practical questions such as privacy and intellectual property to theoretical questions such as the nature of social relations and of humanity itself.

Since this paper attempts to analyse the possibility of computers and information processors as moral agents and will not be able to discuss this question from all possible angles a choice had to be made concerning the ethical theories according to which the analysis is to proceed. The choice here was to concentrate on cognitivist ethics. Briefly, cognitivist ethical theories are understood to be those approaches to ethics which assume that ethical statements can have a truth-value (Höffe, 1992). That rules out all of those approaches which view ethical issues as mere questions of taste or personal preferences. It includes, however, most of the traditional ethical theories which provide us with explicit rules that allow a clear ethical judgment of moral rules or actions. This includes, for example, utilitarian teleology as well as Kantian deontology, the two examples discussed later on. The reason why this group of ethical theories was chosen here is that they have a close affinity to information.

In this view, cognitivist ethical theories are defined by the assumption that some ethical statements can be correctly qualified as true or false. This assumes a worldview in which the truth of statements can be ascertained. It is the same worldview in which information can be true and its truth can be ascertained. This is important because it allows us to draw a connection between ethics and information. Cognitivist ethics assumes that true statements are possible and thereby implies that meaningful and true information is available. It should not be misunderstood as requiring a positivist or objectivist worldview because in order to ascertain the inter-subjective truth of statements non-positivist methods such as Habermasian discourse would also be admissible. The reason why this subset of possible ethical theories was chosen for this topic was that it seems to be the most likely candidate for an ethics which would allow a computer to pass for an autonomous moral agent. In the “objective” world of cognitivist ethics personal qualities are not necessary to be a moral agent.

Within cognitivist ethics there is another relationship between information and ethics. In order for an agent to be able to act or decide ethically she needs information. In order to be able to make ethical decision the agent needs information about the reality of the situation, the norms in question, the expected outcome of actions etc. This is probably true for most non-cognitivist ethics as well but it seems clearest with regard to cognitivist ethics where truth is part of the fundamental normative makeup. Most of the following arguments will aim only at this latter part, at the necessity for an agent to have information in order to be able to make ethical decisions. These arguments are therefore probably valid for other non-cognitivist ethical theories as well. Again, the restriction to cognitivist ethics in this paper was made to strengthen the case for computers as moral agents. In this setting it seems easiest to argue that a computer, in order to act morally, would need to have the right information and apply it correctly. If computers are information processors and can apply information according to higher level rules to information about a situation then it would appear possible to ascribe them the status of moral agents. This chain of reasoning will be developed and critically analysed in the next section.

4. Computers as Autonomous Moral Agents

In this section we will first collect the arguments why computers might appear suitable to play the role of moral agents. For this purpose some thoughts are introduced that seem to suggest that computers could be moral agents. In the second step the problems of this concept are discussed, leading to the third step where the problem of meaning and ethics will be emphasised.

4.1. COMPUTERS AS MORAL AGENTS IN COGNITIVIST ETHICS

It is not possible to discuss the problems of cognitivist ethics in the space of this paper. Even fundamental questions such as whether truth claims in ethical statements have the same nature as truth claims in descriptive propositions (cf. Habermas, 1983, p. 62) will remain open. It will have to suffice for this paper to emphasise the fact that people often talk about moral matters and ethical theories as if differences could be solved by logical arguments. This implies that we believe that ethical statements contain truth of some form. It furthermore implies that we believe in inter-subjective truth and that true information is possible. The idea that information as true and meaningful statements about the world has a relevance for ethics is therefore feasible. For the purposes of this paper it may not even be necessary to settle theses philosophical problems. It may be adequate to point out where computers seem to have strengths and weaknesses regarding the role of information in cognitivist ethics. In order to do so we will briefly look at the role that computers can play in utilitarianism and Kantian deontology.

In order to act morally according to utilitarian theory, one should do what maximises the total utility (Halévy, 1904/1995; Mill, 1976). A computer may do this by simply functioning. However, in this trivial sense, most things might be moral. Even a working water tap would be moral and this use of the concept of morality seems rather doubtful, as will be shown later on. More interesting is the use of computers in order to calculate utility. Among the fundamental problems of utilitarianism we find that it is unclear how utility can be measured and how the aggregate utility can be calculated. At least with the latter problem computers could certainly help given their superior speed of calculation. "[...] it is reasonable to suppose that artificial moral agents, through the wonders of silicon hardware, could take such calculations further than human brains can go" (Allen, 2002, p. 20). Allen does realise that this might lead to a "computational black hole" but he is carefully optimistic that intelligent algorithms might help computers not only calculate utilities but even identify them. Computers could thus develop information necessary to be ethical and process it in such a way that it points the way to moral actions.

From a Kantian point of view the result of an action is not the criterion for its ethical evaluation and computers would need to function differently here. The categorical imperative (Kant, 1995, BA 67) implies that the intention, the maxim, of the agent is of relevance for ethical analysis. The agent needs act to according to a maxim that is universalisable in order to do the right thing. Alan et al. interpret this to mean that "on Kant's view, to build a good AMA would require us to implement certain specific cognitive processes and to make these processes an integral part of the agent's decision-making procedure" (Allen et al., 2000, p. 253). Again, one could find arguments that imply that computers might be made to act ethically in this sense. One can program a computer to follow certain maxims and this could be interpreted to be the computer's equivalent of the "will", which, according to Kant, is what determines the ethical quality of a maxim. Furthermore, the Kantian test of universalisability is at least partly a logical one. A maxim that would make itself impossible through universal use is not universalisable. Kant uses the example of lying, which, in case everybody did it, would be impossible because it would eradicate the truth necessary to lie. Given that computers are based on logic one might hope that they could have a high ability to realise this sort of logical test. Another important aspect of Kantian ethics is impartiality. Kantian ethics claims to be universal and does not allow partial morality. Computers as emotion-free machines might again be said to be in a good position to fulfil this criterion of impartiality.

4.2. SOME PROBLEMS OF COMPUTERS AS MORAL AGENTS

The idea of computers as AMAs produces many problems that cannot be discussed here. First and foremost there is the question whether the definition of ethics allows any other subjects than humans at all. In this paper the position is taken that this is so. However, it remains unclear whether computers can be moral agents. A Kantian, for example, would have to ask whether computers can be autonomous in the Kantian sense (Scanlan, 2000). Autonomy for Kant means the ability to give oneself the laws that one has to follow. It requires freedom and mental faculties that are hard to handle for the philosopher. Another problem for a Kantian may arise from two contradicting maxims, both of which adhere to the Categorical Imperative. How does the computer decide in such a situation? Nevertheless, these questions are almost as complicated to answer with regard to human beings and we will simply neglect them in this paper.

Similarly, there are fundamental problems of utilitarianism that cannot be solved by computers. What is utility, how can it be measured, and how can interpersonal comparisons of utility be realised? If one wants to compare the total utilities of two alternatives in order to make a utility-maximising decision then one needs to know the future, which is impossible. One therefore has to limit the temporal reach of a utility analysis but that is an arbitrary action going counter to the spirit of utilitarianism.

Another problem of ethical approaches which rely on human reason is that the relationship between ethical analysis and moral practice is unclear. Even if a computer could make ethical arguments of an unassailable quality, does that mean that it would be a moral agent? Again, this is a question that in similar form applies to humans and shall be ignored here. For the remainder of the paper we will now concentrate on the question of the relationship of information, ethics, and computers by taking a closer look at the importance of meaning in ethics and the implications this has on the Moral Turing Test.

4.3. MEANING AND COGNITIVIST ETHICS

So far it was argued that information is data with meaning and that information is a necessary precondition for ethics and morality. On this basis it was reasoned that computers as information processors might be autonomous moral agents. In this next section I will argue that the part of information that is relevant to normative questions is its meaning. Computers do not have a sense for the meaning of information and therefore they cannot act as moral agents. The argument will be that computers will fail the Moral Turing Test because they lack an adequate understanding of situations and the possibility to interpret data in a satisfactory manner.

The heart of the argument here is that the moral quality of an action or proposition depends not on any objective criteria but is a social construction that considers the relevant aspects of a situation. In order for a moral actor to participate in this social construct, he or she must understand the meaning of the situation, thus have a complete understanding of the information at hand. This includes the capacity to decide which information is relevant and which is not. My contention is that this is true for all cognitivist ethical theories independent of their theoretical provenance.

Let us look at a simple example: I push an old lady from the sidewalk onto the street. The moral quality of this action depends on the situation and is not determined by the information given so far. For the utilitarian as well as for the Kantian it is important to know whether I pushed the lady because there was a mad skater on the sidewalk, because she asked me to do so, or because I want to kill her by pushing her in front of a lorry. One could argue here that of course nobody has the information about the agent's intention, possibly not even the agent herself. However, other pieces of information may be just as important. An example would be the factual consequences of the action. Does the lady cross the street successfully or does she have an accident? Did a bird fly by in the exact moment and startle me? Did I just take drugs which impeded my judgment? This situation, like every other one, consists of an infinity of facts or data. In order to do an ethical analysis this data has to be ordered, the relevant part must be processed, the irrelevant part must be discarded. There are no algorithms to decide *a priori* which data is relevant and which is not. Human beings are able to make that decision because they are in the situation and they create the relevant reality through interaction. Computers, on the other side, do not know which part of the data they process is relevant for the ethical evaluation because they miss its meaning.

4.4. THE MEANING OF INFORMATION AND THE MORAL TURING TEST

The argument brought forward in the last section might be attacked from several positions. It could be asked what meaning is and why humans can know it or why computers cannot. Metaphysical arguments could be used to demonstrate that it is in fact a circular and anthropocentric argument built on the assumption that only human beings can be moral agents. There may be further hidden assumptions that might be questioned. Such a discussion would be similar the one regarding the question whether computers can display intelligence. Since the Turing Test was introduced to avoid these problems we can now introduce the Moral Turing Test to do the same. If a computer could participate in an interaction with a human being and if it could fool the human interlocutor into believing that it is a moral agent then it could be considered a moral agent.

This approach again has to deal with many possible objections. First, it is doubtful that even a human being might pass the MTT. Behaviour that one observer might consider moral might be thought to be immoral by the next. Different capacities of argumentation might make it difficult for humans to pass the test despite the fact that they are considered moral by their environment. Additionally, there is the conceptual difference between ethical theory and moral practice. All the MTT can do is determining whether a computer could convince an observer of its moral reasoning abilities. Even if it did so, that would say little about the morality of its factual actions or how these could be evaluated.

However, the MTT can be defined as the criterion of moral agency. If a computer passed it then according to this standard it could be considered a AMA. A further narrowing of the MTT to cognitivist ethics would appear to increase the computer's chances by ruling out emotional or intuitive appeals that computers would presumably have difficulties with. Given all these limitations, one could expect computers to pass the MTT. While this is basically an empirical question, it nevertheless seems improbable to me that a computer will succeed in this any time soon. The reason for this pessimism leads us back to the question of information and meaning. In order to competently participate in a dialogue that would allow the observer or "interrogator" to determine whether she is dealing with a moral agent, the computer would need to understand the situation in question. That means that it would have to know the internal states of moral agents, the social process of constructing responsibility ascriptions, and the social background of accepted morality. These three aspects are closely connected and they are only accessible through an understanding of the entire situation.

The point where this argument differs from others such as Allen et al. (2000) or Floridi and Sanders (2001) that are more optimistic regarding computers as AMAs is the location of meaning. Allen et al. assume that computers can make moral statements, which implies that they understand moral meanings. My counterargument is based on Wittgenstein's idea that "the meaning of a word is in its use in the language" (Wittgenstein, 2001, p. 18). In order for a computer to be able to pass the MTT it would have to understand a language and that means it would have to be part of its use and development. This, as far as I can see, is something that computers are not capable of.

There are some aspects regarding this conclusion that should be clarified. First, the argument that computers cannot pass the MTT is not a fundamental one but refers to the current state of computing. In order to develop a principal argument in this direction one would have to show that computers will never understand meaning, something that seems quite difficult. Second, this position would have to concede that in this framework there is little if any difference between the Moral Turing Test and the traditional Turing Test. If a computer passed the latter, there is no good reason to believe it could not pass the former. Third, this last admission returns the argument about the moral status of computers to the well-known field of artificial intelligence. In this framework of cognitivist ethics being an autonomous agent is equivalent to being an autonomous moral agent. Fourth, the argument here says nothing about whether computers should be subjects of moral responsibility. I agree with Floridi and Sanders (2001) that responsibility and accountability ascriptions depend on the level of abstraction. There may be good reasons to hold computers responsible even if they are not autonomous moral agents (cf. Stahl, 2001).

Given these results one could ask whether and under which circumstances computers could become AMAs in the sense that they might be able to pass the MTT.

5. Conclusion: Autonomous Moral Agents the Moral Turing Test

If the argument so far is correct and information is needed for moral agency as determined by the Moral Turing Test and yet computers as information processors are not capable of achieving this then one can try to determine under what circumstances this might change. Or, to put it differently, computers in their current form are not really information processors but only processors of data. The question then is how computers might progress from being data processors to real information processors, how they can access the meaning of the data they process. This question is not trivial and could be asked in a similar way for human beings. How do humans access the meaning of their sensory input?

I believe that the most promising approach to this is the one indicated by Wittgenstein earlier on. It relies on the idea that meaning is not to be found in isolated data but is a social construct that results from interaction. In order for computers to understand information in a meaningful way, which has been argued to be the prerequisite of passing the MTT, they would have to be part of moral discourses. This, however, is problematic because it is a circular requirement. In order to participate in moral discourses one must understand their meaning and in order to understand their meaning one must participate in the discourses. So, how do humans do it? The answer seems to be that humans grow into the position of being moral agents by socialisation, enculturation, and learning.

What would a computer need in order to be able to mimic this development? Part of the answer to this could come from phenomenology. In order to understand meaning one has to be in the situation, to be in the world in a Heideggerian sense (Heidegger, 1993), to share a life-world with others. The life-world is the background of the understanding of meaning, the necessarily shared resource that allows communication, and it constitutes individual meaning as a result of discourses. In order to acquire a life-world an agent needs to be embodied, have emotions, and be able to connect with other participants of discourses on an equal level. Some of these requirements are in fact incorporated in contemporary research in artificial intelligence where the emphasis is shifting from complex and abstract models to embodied interacting agents (cf. Brooks, 2002).

Apart from embodiment a moral agent would need certain cognitive capacities to capture meaning. Most importantly, the agent would need a capacity to learn. The ability to learn combined with embodiment would facilitate reactions to stimuli from the outside world. This allows agents to be socialised into a context of meaning. It also allows them to have a grasp of other agents' life-worlds even if they don't share them. This argument suggests that a top-down programming approach to the acquisition of meaning is not feasible because of the complexity of situations. If computers are to understand meaning and use it adequately then they have to go through a process similar to that which classical moral agents, namely human beings, have to go through.

If these preconditions are given then it might be possible that artificial agents interact with humans and participate in communication. As Apel (1988) and Habermas (1983, 1991) have shown, participation in communication is always a morally charged activity.

In many respects the analysis in this paper has come to the same conclusions that Dreyfus (1993) has spelt out before in his critique of AI. While there is no principal reason why computers will never be able to become moral agents, this attempt to analyse the question with the theoretical lens of information, has shown that there is also no strong hope that this will happen in the foreseeable future. As this paper has tried to show, there are considerable problems with computers as moral agents even if one narrows the question down to cognitivist ethics and if one neglects all of the agency and personhood questions by relying on the Moral Turing Test. But even if computers could overcome these, if they indeed developed an understanding of the meaning of the data they process, the next question would then be whether this would suffice to pass a more general MTT. Maybe emotions, physical and spiritual equality with human beings are necessary for reasoning in a sufficiently human way. The conclusion of this paper is therefore that moral agency of computers is not in sight. While it may be principally possible it is not to be expected soon. The Moral Turing Test is a useful way of moving this debate forward. Nevertheless this paper has argued that even under the most favourable conditions computers are not likely to pass the test. And even if they did, it would require another debate to clarify what this means for their status as moral agents.

References

Allen, C. (2002), 'Calculated Morality: Ethical Computing in the Limit', in I. Smit and G.E. Lasker, eds., *Cognitive, Emotive and Ethical Aspects of Decision Making and Human Action*, Volume I (Workshop Proceedings, Baden-Baden, 31.07.–01.08.2002), pp. 19–23.

Allen, C. et al. (2000), 'Prolegomena to Any Future Artificial Moral Agent', *Journal of Experimental and Theoretical Artificial Intelligence* 12, pp. 251–261.

Apel, K.-O. (1988), *Diskurs und Verantwortung: das Problem des Übergangs zur postkonventionellen Moral*, 3rd edition, 1997, Frankfurt a. M.: Suhrkamp.

Bechtel, W. (1985), 'Attributing Responsibility to Computer Systems', *Metaphilosophy* 16(4), pp. 296–305.

Bloomfeld, B.P. and Coombs, R. (1992), 'Information Technology, Control, and Power: The Centralization and Decentralization Debate Revisited', *Journal of Management Studies* 29(4), pp. 459–484.

Brooks, R. (2002), *Flesh and Machines: How Robots Will Change Us*, New York: Pantheon.

Castells, M. (2000), *The Information Age: Economy, Society, and Culture. Volume I: The Rise of the Network Society*, 2nd edition, Oxford: Blackwell.

Davenport, T.H. and Prusak, L. (1998), *Working Knowledge: How Organizations Manage What They Know*, Boston: Harvard Business School Press.

Dreyfus, H L. (1993), *What Computers Still Can't Do*, Cambridge, MA, London: MIT Press.

Floridi, L. (forthcoming), 'Is Semantic Information Meaningful Data?' in *Philosophy and Phenomenological Research*.

Floridi, L. (2002), 'What Is the Philosophy of Information?' *Metaphilosophy* 33(1/2), pp. 123–145.

Floridi, L. and Sanders J.W. (2001), 'On the Morality of Artificial Agents', in L. Introna and A. Marturano, eds., *Proceedings Computer Ethics: Philosophical Enquiry – IT and the Body*, Lancaster, pp. 84–106.

French, J.A. (1990), *The Business Knowledge Investment: Building Architected Information*, Englewood Cliffs, NJ: Yourdon Press.

Gadamer, H.-G. (1990), *Wahrheit und Methode*, Tübingen: J.C.B. Mohr

Gotterbarn, D. (2002), 'The Ethical Computer Grows Up: Automating Ethical Decisions', in I. Alvarez et al., eds., *The Transformation of Organisations in the Information Age: Social and Ethical Implications*, Proceedings of the sixth ETHICOMP Conference, 13–15 November 2002, Lisbon, Portugal, Lisbon: Universidade Lusiada, pp. 125–141

Grint, K. and Woolgar, S. (1997), *The Machine at Work: Technology, Work, and Organization*, Cambridge: Blackwell.

Habermas, J. (1991), *Erläuterungen zur Diskursethik*, Frankfurt a. M.: Suhrkamp.

Habermas, J. (1983), *Moralbewußtsein und kommunikatives Handeln.*, Frankfurt a. M.: Suhrkamp.

Halévy, E. (1904/1995), *La formation du radicalisme philosophique* (I–III), Paris: Presses universitaires de France.

Heidegger, M. (1993), *Sein und Zeit*, 17th edition, Tübingen: Max Niemeyer.

Höffe, O. (1992), *Lexikon der Ethik*, 4th edition, München: Beck.

Introna, L. (1997), *Management, Information and Power: A narrative of the involved manager*, London: MacMillan.

Johnson, D.G. (2001), *Computer Ethics*, 3rd edition, Upper Saddle River, NJ: Prentice Hall.

Jordan, N. (1963), 'Allocation of Functions Between Man and Machines in Automated Systems', *Journal of Applied Psychology* 47(3), pp. 161–165.

Kant, I. (1995), *Kritik der praktischen Vernunft, Grundlegung zur Metaphysik der Sitten*, Frankfurt a. M.: Suhrkamp Taschenbuch Wissenschaft.

Ladd, J. (2000), 'Ethics and the Computer World — A new challenge for philosophers', in R.M. Baird, R. Ramsower and S.E. Rosenbaum, eds., *Cyberethics — Social and Moral Issues in the Computer Age*., New York: Prometheus Books, pp. 44–55.

Lenk, H. (1994), *Macht und Machbarkeit der Technik*, Stuttgart: Philipp Reclam jun.

Mason, R.O., Mason, F. and Culnan, M.J. (1995), *Ethics of Information Management*, Thousand Oaks, London, New Delhi: SAGE.

Mason, R.O. (1986), 'Four Ethical Issues of the Information Age', *MIS Quarterly* 10, pp. 5–12.

Mill, J.S. (1976), *Der Utilitarismus*, Stuttgart: Reclam Verlag.

Moor, J.H. (2000), 'Toward a Theory of Privacy in the Information Age', in R.M. Baird, R. Ramsower and S.E. Rosenbaum, eds., *Cyberethics — Social and Moral Issues in the Computer Age*, New York: Prometheus Books, pp. 200–212.

Mowbray, M. (2002), 'Ethics for Bots', in I. Smit and G.E. Lasker, eds., *Cognitive, Emotive and Ethical Aspects of Decision Making and Human Action*, Volume I (Workshop Proceedings, Baden-Baden, 31.07.–01.08.2002), pp. 24–28.

Scanlan, M. (2000), 'Does Computer Ethics Compute?' in R.M. Baird, R. Ramsower and S.E. Rosenbaum, eds., *Cyberethics — Social and Moral Issues in the Computer Age*, New York: Prometheus Books, pp. 41–43.

Stahl, B.C. (2001), 'Constructing a Brave New IT-World: Will the Computer Finally Become a Subject of Responsibility?' in R. Hackney and D. Dunn, eds., *Constructing IS Futures* – 11th Annual BIT2001 Conference, Manchester, UK, 30–31 October 2001.

Stahl, B.C. (2000), 'Das kollektive Subjekt der Verantwortung', in *Zeitschrift für Wirtschafts- und Unternehmensethik* 1/2, pp. 225–236.

Stewart, I. (1997), 'Mathematische Unterhaltungen', *Spektrum der Wissenschaft* 7, p. 8.

Stichler, R.N. (1998), 'Ethics in the Information Market', in R.N. Stichler, and R. Hauptman, eds., *Ethics, Information and Technology: Readings*, Jefferson, NC: MacFarland & Company, pp. 169–183.

Turing, A.M. (1950), 'Computing Machinery and Intelligence', *Mind* 59, pp. 433–460.

Ulrich, W. (2001), 'A Philosophical Staircase for Information Systems Definition, Design, and Development', *Journal of Information Technology Theory and Application* 3(3), pp. 55–84.

Wiener, N. (1954), *The Human Use of Human Beings — Cybernetics and Society*, Garden City, NY: Doubleday Anchor Books.

Wittgenstein, L. (2001), *Philosophical Investigations/Philosopische Untersuchungen* (translated by G.E.M. Anscombe), 3rd edition, Oxford: Blackwell.

Zerdick, A. et al., (2001), *European Communication Councel Report: Die Internet-Ökonomie: Strategien für die digitale Wirtschaft*, 3rd edition, Berlin, Heidelberg: Springer