IRIE
International Review of Information Ethics Vol.7 (09/2007)

Michael Nagenborg

Artificial moral agents: an intercultural perspective

Abstract:

In this paper I will argue that artificial moral agents (AMAs) are a fitting subject of intercultural information ethics because of the impact they may have on the relationship between information rich and information poor countries. I will give a limiting definition of AMAs first, and discuss two different types of AMAs with different implications from an intercultural perspective. While AMAs following preset rules might raise concerns about digital imperialism, AMAs being able to adjust to their user's behavior will lead us to the question what makes an AMA "moral"? I will argue that this question does present a good starting point for an intercultural dialogue which might be helpful to overcome the notion of Africa as a mere victim.

Agenda

Author:

Dr. Michael Nagenborg:

- University of Karlsruhe, Institute for Philosophy, Building 20.12, 76128 Karlsruhe, Germany
- ☎ + 49 - 721 – 354 59 55 , ✉ philosophie@michaelnagenborg.de, 💻 www.michaelnagenborg.de
- Relevant publications:
  - Privatheit unter den Rahmenbedingungen der IuK-Technologie, Wiesbaden: VS Verlag 2006.
  - Ethical Regulations on Robotics in Europe, in: AI & Society (in print).

Introduction

At a first glance the concept of „artificial moral agents" (AMA) looks quite spectacular from the perspective of Western philosophy. As I will show in the first paragraph, it's less utopian than one might assume. But the concept still raises serious questions from an intercultural perspective as I will demonstrate in the final paragraph. Since one may ask, if AMAs are a fitting subject for intercultural information ethics, I will point to the relevance of the concept in the context of Africa in the second paragraph.

The purpose of the paper is to show that we need to look at AMAs from an intercultural perspective. Since AMAs are currently used and developed mostly in information rich countries, there is little questioning on their intercultural impact. But since AMAs are designed to follow and enforce moral standards we should be aware that they may cause concern in non-western cultures. They may also be perceived as a tool of the information rich countries which is likely to widen the digital divide between the South and the North.

What is a "artificial moral agent"?

Before asking what is an artificial moral agent, I would like to ask what is an "artificial agent" (AA)? Since I will define an artificial agent first, one might assume correctly, that I do consider AMAs to be a subclass of AAs.

In this paper I will focus on autonomous software agents, although the concept of AMAs is mostly discussed in the context of machine ethics and autonomous robots are a prime example of AAs (Allen et al. 2006). This should also help us to avoid the dangers connected to the humanlike appearance of some robots, which might lead us to accept them as "artificial persons" more easily.

So, what is an "artificial software agent"? One might begin by asking, what is an "agent", but starting with a general definition might again mislead us: Since animals and human beings are considered as "agents", one may think of "artificial agents" as something like "humans" or "animals." Therefore, I will define a "software agent" in contrast to a traditional "software program."

One major difference between a "program" and an "agent" is, that programs are designed as tools to be used by human beings, while "agents" are designed to interact as partners with human beings. I put a special emphasis on "designed … as", because most of the questions, like "is it an agent, or just a program?" (Franklin/Graesser 1996), arise when looking at an existing product. Thus, I suggest that the categories "programs" or "agents" are especially helpful as part of a strategy in software development.[1]

The concept of delegation is a characteristic feature of agents: „An agent is a software thing that knows how to do things that you could probably do yourself if you had the time" (Borking/Van Eck/Siepel 1999: 6). Also, agents may delegate task to other human or artificial agents, or collaborate with other agents. They are designed to perceive the context in which they operate and react to it. Also agents are proactive, therefore one does not need to start an agent (in contrast to a program), but they are designed to decide for themselves when and how to perform a task. Therefore, they may be perceived as autonomous artefacts.

Of course, we have to differentiate between different types of agents according to their capabilities and the degree of autonomy they have. Agents may serve as an interface for human-machine-interaction by acting as an artificial personality, or they might be designed to observe and report on computer systems, aso. What makes the idea of agents interesting with regards to information ethics is that they do raise questions upon the responsibility of the designers as well as the users for the actions carried out by (more or less) autonomous agents. In this paper, however, I will discuss agents on a more general level, since I only want to show that we should have a look at AMA from an intercultural perspective.

From the perspective of Western philosophy one has to be very careful to avoid misunderstanding the concept of "autonomy" in the context of AAs. Surely,

[1] This may be become more obvious by thinking of complex ICT systems which might consist in parts of agents. In the case of internet search engines e. g. web bots might be considered as artificial agents, which are part of a more complex system. This system might also include 'traditional' programs. Does this make a search engine an artificial agent? Although we might ask this question when looking at a specific search engine I assume that such questions do not arise during the design process.

"autonomy" is a central concept at least for the Kantian tradition,[2] but in the context of AA "autonomy" first of all means, that an agent is capable to fulfil a task without direct interference by a human being. One delegates a task to an agent and gets back the results. Here, we should keep in mind the distinction between a "free chooser" and an "autonomous person": A person might be regarded as free, when doing whatever she or he would like to, but we expect an autonomous person also to be someone who thinks about what she or he is doing and does make choices for some reason. I do not want to imply that an autonomous software agent is able to make conscious decisions based on reason, but I do suggest that we expect more than the random results which a free chooser might produce as well. Thus, we expect an artificial agent to fulfil a task while being guided by norms or values. We might expect, e. g., an agent designed to search for scientific literature not to present documents that are obviously not fitting to scientific standards.

Given the explanation of an AA, it is easy to provide a definition of an AMA now: An AMA is an AA guided by norms, which we as human beings consider to have a moral content. To stay with the example of a web bot: One might think of certain content (pornography, propaganda, aso.) as conflicting with moral norms. Thus, an AMA might respect this norms while searching the internet and will not present this kind of content as result unless explicitly being told to do so.

It is important to make a difference between two types of AMAs: Agents may be guided by a set of moral norms, which the agent itself may not change, or they are capable of creating and modifying rules by themselves.[3] But before addressing the two types of AMAs and their different implications, I will ask, why AMAs should be included in the ongoing discussion on intercultural information ethics.

The possible impact of *artificial agents* on Africa

As pointed out by Willy Jackson and Issiaka Mandé (2007: 171):

> *"We have to notice that the ICT are part of all the great issues of globalization … . Unfortunately, we can notice that only a minority take advantage of ICT and thus worsen the inequalities between the rich and the poor, both between the nations and even within the nations. This phenomenon of exclusion and division is particularly visible in the African countries which are the victims of the world economic system."*

There is hope that providing access to ICT and the Internet will provide a link between the information poor and information rich. But as Johannes Britz (2007: 273ff.) has demonstrated there are certain and serious limitations to using the Internet to alleviate information poverty. He pointed out to the importance of physical infrastructures for information societies, the FedEx Factor. Another of these limiting factors is that the content available on the internet is rather useless from the perspective of many non-Western cultures: "… there is indeed more information 'out there', but less meaning" (ibd., 277).

The last point made is important to our subject since artificial agents are designed to help the users to reduce the information overload by filtering and structuring content with regard to the specific needs of the individual users (cf. Kuhlen 1999). Therefore agents are more likely to be used by information rich. This will probably worsen the inequalities between information rich and information poor, since the use of agents may change the nature of the content of the internet. The content will become be less structured according to the needs of human beings, but become more and more accessible to artificial agents. Thus, not having access to agents as mediators to the internet may become a new barrier. Therefore, it is important to keep in mind that changes occurring in information rich countries may indeed have a strong impact on information poor countries.

AAs also may become part of surveillance infrastructures. Here one has to be aware – and this is rather unpleasant to me for being an European author –

[2] The idea of the "autonomy of the practicle reason" is a key feature of Kant's moral theory and is closely linked with the concept of being a person and being able to act according to one's own free will. Autonomy may also be considered to be at the core of human dignity, therefore we should be very careful when applying the concept in the narrow kantian meaning to artificial agents.

[3] Since "autonomous" might be translated as "one who gives oneself its own law", we might assume that not all of this norms are build into the software from the beginning, but the agent is capable of creating new rules for itself.

that already today critics speak of the panoptical fortress of Europe (Davis 2005). As the report on the surveillance society published by the Surveillance Studies Network (2006: 1) points out:

> *It is pointless to talk about surveillance society in the future tense. In all the rich countries of the world everyday life is suffused with surveillance encounters, not merely from dawn to dusk but 24/7.*

Again, the increasing significance of surveillance in rich countries is not restricted to the citizens of these countries but also concerns those intending to (regularly or irregularly) immigrate into these countries (cf. Broeders 2007). Thus, robotic AAs such as the SGR-A1 security system[4] are considered to be only the tip of the iceberg (Rötzer 2006), which should not mislead us to underestimate the importance of AAs with regard to the digital borders limiting the free movement of people as well as information.

But AAs might also provide a better interface for illiterate people, since the idea of speech-based computer-human interaction comes along with the concept of agents as partners.[5] Speech-based AAs serving as interfaces for accessing and creating information might have a great impact on Africa, when considering the already wide spread use of mobile phones.[6] As Rhett Butler (2005) pointed out: "Since computers are rare in much of the region due to poor wire-line infrastructure … and unreliable electrical grids, a technology that offers Internet access without a costly PC promises to pay dividends for Africans." Still, one has also to recognise the results of a case study carried out by Vodafone (2007) that the "use of text messaging in rural communities is much lower due to illiteracy and the many indigenous languages. This has implications for other technologies that use the written word, such as the internet." Thus, providing a speech-based access to the internet by mobile phones might at least provide an opportunity to make more information on Africa by Africans available and accessible to others. Of course, we should not be overoptimistic given what Britz (2007: 274) calls the "Tower of Babel Factor".

Since AAs are designed to lift the weight of dealing with the information overload from the users, they might also help to overcome the "House on Sand factor" (Britz 2007: 277) by enabling users to find relevant information more quickly under the condition that AAs do not need expensive hardware to be used. When AAs are becoming part of online services and may be used in an inexpensive way (or for free), there is also hope that it becomes more and more easy to have access to information needed in a certain context.

I will stop pointing to different issues that may be raised by AAs for now, since the purpose of this paragraph was to demonstrate that AAs are a fitting subject for intercultural information ethics. It is important not to mistake them as too much of "high tech" even when most of the research in this area is carried out in rich countries at the moment, considering the possible positive or negative impact they might have on information poor countries.

What makes an AMA "moral"?

In the first paragraph of this paper I have defined AMAs as a subclass of artificial agents that include what Colin et al. (2006: 14) have called an "ethical subroutine". Further, I suggested to differentiate between AMAs that are guided by moral norms, which they are can't change, and AMAs that may produce moral norms by themselves.

AMAs not being able to change their "ethical subroutine" are autonomous in the action they take, but they are not able to do "bad things". A good example of such an AMA is the main character of the movie "RoboCop" (USA 1987), who is not capable to overcome the prime directives which he was programmed to follow. But search engines such as Google might be considered to be AMAs of this type as well, if we agree that they are Aas, too. At least, such search engines may be regarded as autonomous systems, since the results they produce may not be foreseen neither by the software developers nor the users. And especially services such as

[4] < http://www.samsungtechwin.com/product/features/dep/SSsystem_e/SSsystem.html > (retrieved on July 8, 2007)

[5] One might think of the digital butler described by Negroponte (1995) as a good example of this type of AA.

[6] When thinking about speech-based human-computer-interaction one should keep in mind that – to use the WSIS wording – the right to communicate does include the right to read and the right to write.

"Google Alerts"[7] may be considered as AAs since they act without direct control of their human users. There might be arguing that these are very simple services, but we are not concerned with the level of autonomy here. What is more important is that they are autonomous and that they are – at least in Germany – limited by norms, which are considered to be moral norms. As already stated above it might be considered as a moral norm that no documents that may be harmful for children (like pornography, excessive depictions of violence, and hate speech) are presented to children. Thus, the German law does not allow to make this kind of documents available to persons under the age of 18, and also bans the distribution of certain documents at all. Now, their was some concern that these kind of online services undermine these legal standards (cf. Neuberger 2005), which lead to a voluntary agreement signed by all major search engines on not providing links to German users which point to documents banned in any other kind of media. Therefore, at least the German versions of these search engines might be regarded as AMAs, since they include services to be considered as AAs and they are limited by "ethical subroutines".

The question if such kind of censorship may be considered ethical is less important from an intercultural perspective than the question of the impact such AMAs may have on other cultures. Even without AAs on the Internet, there has been questioning about the values embedded unconsciously in computer-mediated communication by their Western designers (Ess 2007: 153). Thus, there must be questioning about what kind of "morality" will be fostered by AMAs, especially since now norms and values are to be embedded consciously into the "ethical subroutines". Will they be guided by "universal values", or will they be guided by specific Western or African concepts? Maybe, the kind of filtering in accordance with the German Law might be acceptable and even desirable from an African perspective. But how about AMAs designed to protect privacy? There are already first steps taken in the development of such AMAs which are also presented as an example in the context of machine ethics (Wallach et al. 2006: 13). What would be the impact of such AMAs on cultures, which are characterised by a more community based thinking and therefore do not value privacy in the same way as Western cultures (cf. Olinger et al. 2005)?

The second type of AMAs being able to create rules of behaviour by themselves for themselves in accordance to their users' preferences might be seen as an alternative in this perspective for they should be able to adjust to the specific cultural background of their users. Such an agent could learn, i. e., what kind of norms are followed by an European or an African user. Beside of the question how to deal with "bad users" training the AMAs to behave unethically, there should be questioning on what are the distinctive features of a moral norm and what makes such a norm different from, i. e., legal norms? And what should an agent do when it is given a task that an user finds to be legitimate and even necessary from the moral point of view, but is conflicting with legal norms?

The challenges arising from such questioning are not only to be considered pragmatically, but are also a good starting point on an intercultural dialogue on AMAs, which goes beyond the notion of "digital imperialism", an issue that might be raised with regards to the first type of AMAs presented above. That is not to say that digital imperialism is not to be regarded as an ethical issue; but thinking of the requirements of an AMA has to fulfil to be regarded as "morally" (in the limited sense introduced in the first paragraph) does offer a opportunity to go beyond the idea of Africa being a mere victim of Western technology. Rather, it will enable us to discuss the rich offers in African thinking on what it means to be an autonomous moral agent (cf. Sogolo 1993: 129ff) by asking what we are going to expect from AMAs and which is truly a moral agent and not just a learning agent.

References

Broeders, D. (2007): The New Digital Borders of Europe. EU Databases and the Surveillance of Irregular Migrants. In: International Sociology 2007, Vol. 22 (1), 71-82.

Butler, R. (2005): Cell phones may help "save" Africa. Online: < http://news.mongabay.com/2005/0712-rhett_butler.html > (retrieved on July 8, 2007)

Capurro, R., J. Frühbauer, T. Hausmanninger (Eds.): Localizing the Internet. Ethical aspects in intercultural perspective. München: Wilhelm Fink 2007.

Colin, A., W. Wallach, I. Smit (2006): Why Machine Ethics? In: IEEE Intelligent Systems, Vol. 21, No. 4, pp. 12-17.

[7] < http://www.google.de/alerts?hl=eng > (retrieved on July 8, 2007)

Davis, M. (2005): The Great Wall of Capital. In: M. Sorkin (Ed.): Against the Wall. New York – London: The New Press, pp. 88-99.

Ess, C. (2007): Can the Local Reshape the Global? Ethical Imperatives for Humane Intercultural Communication Online. In: Capurro/Frühbauer/Hausmanniger 2007, pp. 153-169.

Franklin, St., and A. Graesser (1996): Is it an Agent, or just a Program? < http://www.msci.memphis.edu/~franklin/AgentProg.html > (retrieved on July 8, 2007)

Jackson, W., and I. Mandé (2007): "New Technologies" and "Ancient Africa": The Impact of Information and Communication Technologies in Sub-Saharian Africa. In: Capurro/Frühbauer/Hausmanniger 2007, pp.171-176.

Kuhlen, R. (1999): Die Konsequenzen von Informationsassistenten. Frankfurt am Main: Suhrkamp.

Negroponte, N. (1995): Being digital. New York: Alfred A. Knopf.

Neuberger, C. (2005): Function, Problems, and Regulation of Search Engines in the Internet (Extended Abstract). In: International Review of Information Ethics, Vol. 3 (06/2005). < http://www.i-r-i-e.net/inhalt/003/003_neuberger_ext.pdf > (retrieved on July 8, 2007)

Olinger, H. N., J. J. Britz, M. S. Olivier (2005): Western privacy and ubuntu: influences in the forthcoming data privacy bill. In: P. Brey, F. Grodzinsky, L. Introna (Eds.): Ethics of New Information Technology. Proceedings of the Sixth International Conference of Computer Ethics: Philosophical Enquiry (CEPE2005). Enschede, NL: CEPTES, pp. 291-306.

Rötzer, F. (2006): Kampfroboter zum Schutz von Grenzen, Flughäfen oder Pipelines. Online: < http://www.heise.de/tp/r4/artikel/23/23972/1.html > (in German) (retrieved on July 8, 2007)

Sogolo, G. (1993): Foundations of African Philosophy. Ibadan: Ibadan University Press.

Vodafone (2007): Impact of Mobile Phones in Africa. Online: < http://www.vodafone.com/start/responsibility/our_social___economic/socio-economic_impact/impact_of_mobile_phones.html > (retrieved on July 8, 2007)