Artificial moral agents: An intercultural perspective

Michael Nagenborg

This chapter argues 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. A limiting definition of AMAs is given first, followed by a discussion of 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 users' behaviour 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 that might be helpful to overcome the notion of Africa as a mere victim.

Contents

Author's details

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

Introduction

At a first glance, the concept of "artificial moral agents" (AMAs) looks quite spectacular from the perspective of Western philosophy. As I will show in the first section, it is less utopian than one might assume. But the concept still raises serious questions from an intercultural perspective, as I will demonstrate in the final section. Since one may ask whether 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 section.

The purpose of the chapter 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. However, 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 an "artificial moral agent"?

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

In this chapter 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 "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 programme".

One major difference between a "programme" and an "agent" is that programmes 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 the words "designed as", because most of the questions (like: "Is it an agent, or just a programme?"; Franklin & Graesser, 1996), arise when looking at an existing product. Thus, I suggest that the categories "programmes" 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 et al., 1999:6). Also, agents may delegate tasks 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 programme), 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. What makes the idea of agents interesting with regard to information ethics is that they do raise questions about the responsibility of the designers, as well as the users, for the actions carried out by (more or less) autonomous agents. In this chapter, however, I will discuss agents on a more general level, since I only want to show that we should have a look at AMAs from an intercultural perspective.

[1] This may become more obvious by thinking of complex information and communication technology (ICT) systems that might consist in parts of agents. In the case of Internet search engines, web bots for example might be considered artificial agents, which are part of a more complex system. This system might also include "traditional" programmes. 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.

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, "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 makes 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 that a free chooser might produce as well. Thus, we expect an artificial agent to fulfil a task while being guided by norms or values. For example, we might expect an agent designed to search for scientific literature not to present documents that obviously do not fit scientific standards.

Given the explanation of an AA, it is now easy to provide a definition of an AMA. 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, etc.) as conflicting with moral norms. Thus, an AMA might respect these norms while searching the Internet and will not present this kind of content as a result, unless explicitly 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 that the agent itself may not change, or they are capable of creating and modifying rules by themselves.[3] However, 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 Jackson & Mandé (2007:171):

> *We have to notice that the ICTs 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 points 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".

The last point made is important to our subject, since artificial agents are designed to help users to reduce 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 the 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. 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. It is therefore important to keep in mind that changes occurring

[2] The idea of the "autonomy of the practical 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 these norms are built into the software from the beginning, but the agent is capable of creating new rules for itself.

in information-rich countries may indeed have a strong impact on information-poor countries.

AAs may also become part of surveillance infrastructures. Here one has to be aware – and this is rather unpleasant to me as a European author – that critics are already speaking 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 immigrate (regularly or irregularly) 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). This 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.

AAs might also, however, 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 widespread use of mobile phones.[6] As Butler (2005) points 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 to recognise the results of a case study carried out by Vodafone (2007), showing 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 speech-based access to the Internet through 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".

As AAs are designed to lift the weight of dealing with 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 free of charge), there is also hope that it would become 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 about AAs for now, as the purpose of this section was to demonstrate that AAs are a fitting subject for intercultural information ethics. It is important not to mistake them for being too "high tech", even when most of the current research carried out in rich countries, considering the possible positive or negative impact they might have on information-poor countries.

What makes an AMA "moral"?

The first section of this chapter defined AMAs as a subclass of artificial agents that include what Allen et al. (2006:14) have called an "ethical subroutine". Further, I have suggested that one should differentiate between AMAs that are guided by moral norms, which they cannot change, and AMAs that may produce moral norms by themselves.

AMAs that are not 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 US movie "RoboCop" (1987),

[4] http://www.samsungtechwin.com/product/features/dep/SSsystem_e/SSsystem.html. Accessed 8 July 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 wording of the WSIS – the right to communicate does include the right to read and the right to write.

who is incapable of overriding the prime directives that he has been programmed to follow. Search engines like 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, as the results they produce may not be foreseen either by the software developers or the users.

In particular, services such as "Google Alerts"[7] may be considered AAs because they act without direct control of their human users. One could argue 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 – at least in Germany – limited by norms that are considered moral norms. As stated above, it might be considered a moral norm that no documents that may be harmful (like pornography, excessive depictions of violence, and hate speech) are presented to children. Thus, German law does not allow making these kinds of documents available to persons under the age of 18, and also bans the distribution of certain documents. There has been some concern that these kinds of online services undermine such legal standards (cf. Neuberger, 2005), which has led to a voluntary agreement being signed by all major search engines to not provide links to German users that 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 AAs and are limited by "ethical subroutines".

The question whether such a 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 have been questions about the values embedded unconsciously in computer-mediated communication by their Western designers (Ess, 2007: 153). Thus, thought must be given to what kind of "morality" will be fostered by AMAs, especially since norms and values are now to be embedded consciously into the "ethical subroutines". Will these be guided by "universal values", or by specific Western or African concepts? Maybe, the kind of filtering done in accordance with German law might be acceptable and even desirable from an African perspective. But what about AMAs designed to protect privacy? Already, first steps have been taken in developing such AMAs, which are also presented as an example in the context of machine ethics (Allen et al., 2006:13). What would the impact of such AMAs be on cultures that are characterised by a more community-based thinking, and therefore do not value privacy in the same way as Western cultures do? (Cf. Olinger et al., 2005.)

The second type of AMAs that are able to create rules of behaviour by themselves for themselves in accordance with 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, for example, what kinds of norms are followed by a European or an African user. Besides the question of how to deal with "bad users" training the AMAs to behave unethically, there should be discussion on what the distinctive features of a moral norm are and what makes such norms different from, for example, legal norms. Moreover, what should an agent do when it is given a task that a user deems legitimate and even necessary from a moral point of view, but is conflicting with legal norms?

The challenges arising from such questions are not only to be considered pragmatically, but are also a good starting point for an intercultural dialogue on AMAs that goes beyond the notion of "digital imperialism", an issue that might be raised with regard to the first type of AMAs presented above. That is not to say that digital imperialism is not to be seen as an ethical issue, but thinking of the requirements that an AMA has to fulfil to be regarded as "moral" (in the limited sense introduced in the first section) does offer an 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 are truly moral agents and not just learning agents.

[7] http://www.google.de/alerts?hl=eng. Accessed 8 July 2007.

REFERENCES

Allen, C., Smit, I. & Wallach, W. 2005. Artificial morality: Top-down, bottom-up and hybrid approaches. *Ethics and Information Technology*, 7: 149–155.

Allen, C., Wallach, W. & Smit, I. 2006. Why machine ethics? *IEEE Intelligent Systems*, 21(4): 12–17.

Borking, J.J., Van Eck, B.M.A. & Siepel, P. 1999. *Intelligent software agents and privacy.* Publication of the Dutch Data Protection Authority, The Hague.

Britz, J. 2007. *Critical analysis of information poverty from a social justice perspective*. D.Phil. dissertation. Information Science School of Information Technology, University of Pretoria, South Africa.

Broeders, D. 2007. The new digital borders of Europe: EU databases and the surveillance of irregular migrants. *International Sociology,* 22(1): 71–82.

Butler, R. 2005. *Cell phones may help "save" Africa.* http://news.mongabay.com/2005/0712-rhett_butler.html. Accessed 8 July 2007.

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

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

Ess, C. 2007. Can the local reshape the global? Ethical imperatives for humane intercultural communication online. In Capurro, R., Frühbauer, J. & Hausmanninger, T. (Eds), *Localizing the Internet: Ethical aspects in intercultural perspective.* München: Wilhelm Fink, 153–169.

Franklin, S. & Graesser, A. 1996. *Is it an agent, or just a program?* http://www.msci.memphis.edu/~franklin/AgentProg.html. Accessed 8 July 2007.

Jackson, W. & Mandé, I. 2007. "New technologies" and "Ancient Africa": The impact of information and communication technologies in sub-Saharan Africa. In Capurro, R., Frühbauer, J. & Hausmanninger, T. (Eds), *Localizing the Internet: Ethical aspects in intercultural perspective.* München: Wilhelm Fink, 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. *International Review of Information Ethics,* 3(6). http://www.i-r-i-e.net/inhalt/003/003_neuberger_ext.pdf. Accessed 8 July 2007.

Olinger, H.N., Britz, J.J. & Olivier, M.S. 2005. Western privacy and *ubuntu*: Influences in the forthcoming Data Privacy Bill. In Brey, P., Grodzinsky, F. & Introna, L. (Eds), *Ethics of new information technology.* Proceedings of the Sixth International Conference of Computer Ethics, Philosophical Enquiry (CEPE2005). Enschede, NL: CEPTES, 291–306.

Rötzer, F. 2006. *Kampfroboter zum Schutz von Grenzen, Flughäfen oder Pipelines.* http://www.heise.de/tp/r4/artikel/23/23972/1.html. Accessed 8 July 2007.

Sogolo, G. 1993. *Foundations of African philosophy.* Ibadan: Ibadan University Press.

Surveillance Studies Network. 2006. *A report on the surveillance society.* http://www.ico.gov.uk/upload/documents/library/data_protection/practical_application/surveillance_society_full_report_2006.pdf.

Vodafone. 2007. *Impact of mobile phones in Africa.* http://www.vodafone.com/start/responsibility/our_social_economic/socio-economic_impact/impact_of_mobile_phones.html. Accessed 8 July 2007.