*Ethics in Progress* (ISSN 2084-9257). Vol. 10 (2019). No. 2,
Art. #6, pp. 52-67. DOI:10.14746/eip.2019.2.6

# *The Ethical Significance of Human Likeness in Robotics and AI*

Peter Remmers

(Technical University of Berlin; peter.remmers@tu-berlin.de)
ORCID: 0000-0001-5625-3144

## Introduction

Thinking about the complex relations between minds and machines, some apparently diverging phenomena suggest a vaguely two-faced character of certain technologies. To give a few examples in a polemical spirit: Ever since its inception, the research paradigm of artificial intelligence (AI) generated computer programs that enable us to have conversations with machines in a way very much like conversations with other human beings. But while Joseph Weizenbaum was concerned about the tendency of laypersons to use simple computer programs as professional psychoanalysts, researchers in human-robot interaction utilize similar behavioral tendencies to advance synergies between machines and persons. In a related endeavor, robots like Ishiguro's "geminoids" or Hanson Robotic's "Sophia" are made to look very similar to humans. But while some scholars dismiss these robots as superficial marketing curiosities with no intellectual potential, others see in them a canny light on the horizon of the wintry valley. Finally, people worry about robots and AI systems becoming so versatile and knowledgeable that automation will finally free them from needless labor. But while the ultimate goal of robots to replace, destroy or enslave the human race has now become an old meme, people involved in actual human-robot interactions occasionally find themselves speculating about what the robot might think about them just now.

The seemingly divergent perspectives marked in these examples can be interpreted in terms of certain tendencies in AI and robotics. On the one hand, a defining goal of AI and robotics is to build technical artifacts as substitutes, assistants or enhancements of human decision-making and action. For without the aspiration to create tools for automation, optimization and interaction, AI and robotics would not be the significant technologies they are. On the other hand, we sometimes encounter certain kinds of human

likenesses in artifacts while interacting with these technologies – and when we do, it is not always clear what to make of it. How far does the likeness between technical artifacts and humans go? Is it a mandatory part of AI and robotics? Can there be kinds of human likeness in technical artifacts that call for an adaption of our behavior towards them? As we will see in the following sections, the issues involved might range from harmless cultural references to significant ethical issues.

In the present contribution, I want to highlight three aspects emerging from human likeness in AI and robotics. First, I will broadly investigate some of the reasons that motivate the exclusion of the philosophical debate on artificial minds from AI research. As approaches to AI and robotics are themselves not strictly guided by models of human performance, it is generally not their objective to reproduce human cognition and action. Still, there are some unavoidable relations between the ideas of human likeness and artificial agents. To clarify these relations, I will briefly recall and arrange certain well-known points from related debates in a guided tour through seemingly different topics in AI and robotics.

Closely connected is the second aspect to be discussed, namely the question if human likeness in technical artifacts of AI and robotics might lead to some ethical confusion about their moral status. In the course of the discussion, I will argue for the claim that although it might seem that there are ethical issues involved in certain philosophical speculations, methodical operationalizations or terminological ambiguities of AI and robotics, some underlying confusions are easy to clarify and explain (however persistent in non-professional discourse).

Third, I will suggest in the last section that although human likeness may not be ethically significant on the philosophical and conceptual levels, strategies to use *anthropomorphism* in the technological design of human-machine interactions are ethically relevant. Here, issues arise because specific kinds of relations between humans and machines are developed. Finally, it should become clear that the really important ethical issues are concerned with interactive scenarios involving actual technical artifacts, or, more generally, with technology and its design.

Accordingly, in the following sections, I will move from theoretical issues as discussed in philosophical debates on strong AI to ethical issues arising from strategies of anthropomorphism in human-machine interaction. In the next section, the philosophical debate on 'strong AI' is very briefly interpreted in the context of the philosophy of mind, whereas its relevance for technological research is questioned on pragmatic grounds. The 3rd section deals with an influential strategy to sideline philosophical implications of human likeness in AI research, exemplified by the famous Turing Test. In the 4th section, the ambiguous concept of autonomy serves as an example for unintentional anthropomorphism on the conceptual level. Finally, in the last section, intentional anthropomorphism is presented as a design strategy to support collaborations in human-machine interaction. In the end, some of the different roles and functions of human likeness

in AI and robotics and their implications for the ethical status of technical artifacts should present themselves clearly.

## 1. The Ontological Issue of Strong AI: Can Machines Think?

When we apply concepts of mind to technical artifacts, we often speak metaphorically. This is especially true for our interactions with robots and AI systems. Although we might call the performance of a computer program 'clever', the facial expressions of a robot 'friendly' or the defective movements of a mobile robot 'moody', we know that technical artifacts are not *really* clever, friendly or moody. We are well aware that attributions like these are traditionally used to refer to living beings, mostly humans, and we usually hold on to the difference between human persons and nonliving artifacts – at least, for now, in practice. But while there is no dissent about the fact that we often talk about *present* technology by using metaphors derived from human domains, there seems to be no consensus about *future* possibilities. The question comes up: given the rapid and overwhelming advancement of technology in robotics & AI, will the metaphorical ascription of human attributes at some point become non-metaphorical?

The idea of a non-metaphorical way to address artificial systems by using familiar human attributes and concepts is a well-known topic in Science Fiction. In philosophy, we are thinking in a similar spirit about general concepts like 'mind', 'intelligence', 'autonomy' or 'social relations', asking questions like the following: Can we apply the concept of mind to artificial systems? How does artificial intelligence relate to human thinking? And if artificial agents can have a mind, shouldn't we be prepared to treat them as entities with some kind of moral status? By asking these questions, philosophical and ethical implications of deeply entrenched concepts and fundamental practices are at stake. They mark the transition from metaphorical, often playful idioms to serious issues of emerging technologies. We refer to technical artifacts that would be the objects of non-metaphorical ascription of human attributes as 'strong AI'. The vision of strong AI is embodied in the technological aim to produce artificial minds, i.e. non-human, technologically created minds.

To summarize, the basic question is: Is strong AI possible? The issue at stake is an *ontological* one. When we ask ontological questions, we are asking about the way something *is* – as opposed to *epistemological* questions about how we come to know about something, or as opposed to *phenomenological* questions about how something appears to us. Consequently, we could phrase the ontological question in this way: Is it indeed possible for certain artificial systems like robots or AI agents to *think* or to *have a mind*, as opposed to the uncontroversial possibility of *simulating* thinking or *simulating* a mind? To label this question 'ontological' means that we do not ask about empirical evidence for the presence of thinking or of a mind. Instead, we ask about what 'intelligence', 'thinking' and 'mind' actually are and if the idea to create an artificial mind is coherent at all. In

other words: What is the 'nature' of mind and what would it mean to make one?[1] With questions like these, we express a philosophical interest in the ontological status of mind.

The ontological discourse about artificial minds is usually the subject of thought experiments, closely connected to common themes in science fiction. But some answers to these theoretical questions might turn out to be helpful for an answer to an even more thrilling follow-up question to the idea of strong AI: "Is it *technologically* possible for us to build artificial agents that actually have a mind?" If a philosophical theory of mind would be able to provide a glimpse of an answer, this would constitute a direct contribution of philosophy to technology, and for this very reason some philosophers are extraordinarily excited about the topic.

Finally, the ontological question might have *ethical* implications. If the question would be confirmed – i.e. if artificial agents actually *were* intelligent as opposed to simulating intelligent performance – this fact might provide reasons for us to treat robots as we would treat any intelligent being, i.e. humans or other living beings. Whatever the grounds for the moral status of persons are, it certainly seems to have something to do with the fact that persons have a mind. By consequence, artificial agents would have to be considered if it turned out they really have a mind, too. The ethical consequences of this reasoning are well-known from many Science Fiction stories.

Now could an artificial realization of cognitive processes also have (or, more precisely, *be*) a 'mind'? For the purpose of the present contribution, I do not intend to take a stand in the ontological debate about strong AI or artificial minds; nor will I discuss the involved positions and arguments. Instead, without addressing the question directly, I want to highlight a strategical observation about the ontological question that provides a reason to push it into the background of AI research.

To begin, it is not entirely irrelevant to remember the philosophical origins of the debate. Although the idea of artificial living beings goes back to ancient myths, the present debate originates from the philosophy of mind in early modern thinking. Bluntly summarized, revolutions in the emerging empirical sciences and in technology inspired metaphysical doctrines of materialism, including, more specifically, the idea of living beings as mechanical systems. Following mechanism, more sophisticated doctrines like physicalism, causalism or functionalism branched out in the long and winding philosophical debate. In addition to historically earlier stages of the debate, the contemporary setting is influenced by the interpretation of 'intelligence' in terms of cognitive processes (i.e. mental states and processes) in combination with formal theories of information and computation as developed in the middle of the twentieth century. In summary, many debates in the philosophy of mind revolve around the relation between the mind and the general mechanical / physical / causal / functional / computational / informational

1 Obviously, these are two different questions. But they are often taken to be closely related, as exemplified by the title of a paper of Fred Dretske about intentionality: "If You Can't Make One, You Don't Know How It Works" (Dretske 1994).

principles of the world. Consequently, the line of the 'materialist' argument might take the following form: for as long as we do not find good reasons for the exclusion of human minds from the scientifically accessible features of the world, it is reasonable to assume that the human mind can basically be described according to these principles. And if we admit that the mind depends on some of these principles, there is no fundamental reason to doubt the possibility of reproducing it artificially.[2]

According to this very rough sketch, AI research carries strong philosophical and anthropological presuppositions on its back, as discussed in philosophical debates on computers and especially on the possibility of strong AI. But just like in every emancipated scientific discipline, the philosophical background does not interfere with AI research; its ties to philosophical and anthropological concerns were cut early in its history, as described in the next section. Consequently, the ontological question of whether strong AI is possible is only of philosophical interest at best.

This separation of AI research from its philosophical issues is strengthened when we consider the role technical artifacts are supposed to play in the context of the ontological question: On the one hand, the technological development of strong AI in the described sense would obviously prove its theoretical possibility. On the other hand, even if it might be *theoretically* possible to build a machine that would be generally accepted as being *actually* intelligent, it seems unclear what *practical* value it would add to any machine that only *simulated* its intelligent performance (while performing identically in every other respect). Granted that the research and development to achieve strong AI would be very expensive and time-consuming, one would expect a clearly defined practical vision to justify the investment of resources. Given certain alternatives as outlined in the next section, the development of strong AI does not seem to be cost-effective, simply because there is no apparent and convincing practical utility for it – besides proving a few philosophical claims.[3]

Even if this argument against the significance of strong AI is pragmatic in nature,

2 Although materialism is the generally accepted paradigm, opposition to materialist positions is manifold. It is sometimes motivated by anthropological or ethical concerns. For example, some authors argue that the idea of minds as machines contradicts the nature and/or value of humanity (Weizenbaum 1976; Lanier 2010). In the ensuing debate, materialists usually argue against any negative ethical consequences of their materialist doctrine. Unfortunately, the issue is sometimes framed by interpreting concerned positions as instances of unscientific intellectual narcissism, for example by discussing them under the header "People might lose their sense of being unique" (Russell & Norvig 2016, 1035).

3 One might expect a use case of strong AI in the contemporary vision to develop technological ways of 'copying' an individual's mind to a data storage ('mind uploading'). The idea inspires theoretical issues about the mind and the brain, but according to some neuroscientists, its technological realization is not very close: „It will almost certainly be a very long time before we can hope to preserve a brain in sufficient detail and for sufficient time that some civilization much farther in the future, perhaps thousands or even millions of years from now, might have the technological capacity to 'upload' and recreate that individual's mind" (Miller 2015). (But note that the company *Carboncopies* actively pursues the plan to reach this goal within short time.) Be that as it may, the vision of copying minds to data storage depends on speculative predictions, while the theoretical arguments for and against its possibility are discussed independently from the actual technological progress.

it certainly limits the prospects for any philosophical contribution to technological outcomes. We might conclude that philosophy is – again – thrown back to its original purpose, namely to think about very basic and general issues in order to clarify them, but without the prospect of actively contributing to technological development. We might go even further and take this pragmatic outlook as a reason for some general skepticism about research in AI.[4] But this skepticism should be limited to bold claims about the technological achievement of strong AI. As a result, a more realistic approach to contemporary research in AI is called for. AI and robotics are indeed originally inspired by some features of humans and other living beings, but the actual technological developments take a different route in most respects, as briefly outlined in the next section. Therefore, while ethical implications concerning our behavior towards machines might result from philosophical speculations on strong AI, they are nothing to *actually* worry about – the technology is detached from the philosophical prospects.

## 2. Operationalizing Intelligence: Can Machines Act Intelligently?

When we think of actual technological approaches, the idea of strong AI does not seem relevant for AI research at all. Instead, AI research aims at an optimization or automation of certain actions and performances we find in human practice. This approach leads us away from the idea of an original 'mind' in a machine towards the idea of 'intelligent behavior'. According to this strategy, we are trying to find a general model of intelligent behavior, i.e. a description on a general level that covers intelligent performance by humans as special cases. This approach is exemplified by Alan Turing's famous Test.

In his influential article presenting the Turing Test, Turing initially seems to be preoccupied with an ontological question as discussed in the previous section: "I propose to consider the question, 'Can machines think?'" (Turing 1950/2004, 441). But he immediately goes on to express his indifference to any semantical or philosophical debate:

> This should begin with definitions of the meaning of the terms 'machine' and 'think'. (...) Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words (Turing 1950/2004, 441).

The new question refers to the famous 'imitation game', an experimental setup devised in order to isolate intelligent performance from its 'mental' origins: Would

4 Although the present argument contends that strong AI is not likely to happen, it should not be understood as a kind of "argument from disability", claiming that AI systems "can never do X" (Russel & Norvig 2016, 1021). The point is not that AI systems would be unable to do something. Instead, the point is that their ability to do anything is not altered by the ontological state of the involved cognitive processes, i.e. by an artificial agent's ability being *real* or just *simulated*.

a human interrogator be able to tell the difference between a human and a computer program only based on an examination of their intelligent performance?

But the goal of the Turing Test is not an imitation of human intelligence itself; imitation is merely used as a way to measure intelligence without any reference to mental states. To understand this interpretation, we have to keep in mind that at the time of Turing's writing, behaviorism was the leading paradigm of psychology. According to behaviorism, what we call 'mind' is fully explained by references to behavior. On this ground, further questions about the nature of mind or about mental states are dismissed. Consequently, if a machine exhibits intelligent behavior, it *can* think. From the perspective of behaviorism, there is no room for further speculation on the 'inner' (i.e. unobservable) workings of intelligence (or the 'mind') beyond the observable behavior.

Shortly after Turing's influential article, cognitivism emerged as a contender of behaviorism. As a psychological paradigm, cognitivism allows and demands inferences from behavioral data to processes located 'in the mind' (or, as in cognitive neuroscience, in the brain). According to the cognitive paradigm, it is perfectly reasonable to ask if a machine can *actually* think or if it just performs *as if* it could think. Both operations would originate from different cognitive processes. The cognitivist paradigm partly explains some criticism of the Turing Test, most prominently represented by the famous thought experiment of the 'Chinese Room' by John Searle (Searle 1980). Searles argument is directed against the position of 'strong AI', construed as the position that "the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states." (Searle 1980, 417) To followers of a strictly behaviorist paradigm, Searle's argument against strong AI comes to nothing, while from a cognitivist perspective, it creates an open question not answerable by Turing's Test – at least if we take Turing's Test as an approach to answer the ontological question of strong AI's possibility.

But the Turing Test is not a merely historical piece of psychological theory, as outdated as behaviorism itself – quite the contrary. What Turing achieves with the conception of the Turing Test should be more accurately described as a step towards an operationalization of 'intelligence'. And this direction of operationalization is an essential part of any research in AI since Turing. Here, we are not interested in the ontological difference between real intelligence and simulated intelligence; what counts is the outcome, the performed action. The aim of the computer's imitation of human behavior is not human likeness as such – human intelligence just works as a measure. In principle, any intelligent being could take the place of the human being in the test – the fact that only humans perform intelligently is taken to be purely accidental. The measure of intelligence then works as a reference for the technological objective to create an intelligent artifact we can actually make use of in practice, i.e. automation of intelligent performance. To achieve this, we need a technological operationalization of the task at hand. Leading questions then are: What information can we make available to a machine, what is it supposed to

do with it, and what is the desired outcome? An engineer has to come up with answers to these questions, formulated in a way he can make use of in research and development.

Still, there is a methodical issue in the original Turing Test, as it is not focused on a sufficiently task-specific intelligent performance. In its original form it is restricted to communication by written language, but it broadly involves any kind of communicative behavior we might expect from humans. In contrast to Turing's original setup, contemporary developments in AI technology are usually focused on more specifically defined skills. Following this tendency, current methods of machine learning have already shown some surprising and impressive results. The process of formulating algorithms for machines that describe the steps to accomplish a certain task is a way of isolating basic elements of intelligent performance. From the perspective of human intelligence we generalize or abstract the relevant parts of human perceptions, plans and actions from their 'original' occurrence in human practice. In an effective automatization of a task, the information involved in the process will present itself differently to a machine and to a human. The technological formulation of algorithms, taken literally, would probably not be useful for a human trying to achieve the same task. What humans need to accomplish a task is very different from what a machine needs to accomplish the same task. Thus, any approximation of cognitive processes to mental processes is not a viable strategy in this context.

For this reason, actual technological developments in robotics and AI are *not* modelled after the human mind; at best, they work with abstractions and operationalizations of intelligent performance to arrive at an effective automation of some specific tasks. For example, the concept of an 'intelligent agent' is a generalization of human agency in functional terms. While humans perceive, think and act, an intelligent agent is more generally described as an entity that has an input, a processing unit and an output. Humans are intelligent agents whose input is specified by their sense organs, whose processing unit is identified with their brains (or their bodies) and whose output consists in their actions. When we think about the concepts of mind and intelligence, we can say that human and artificial intelligence will generally be different in this way (at least).

## 3. The Pitfalls of Ambiguous Terminology: Technological vs. Personal Autonomy

This interpretation of AI research provides a clear distinction between intelligence as present in human behavior and general accounts of intelligence as applied to artificial agents. As a consequence of this operationalization, the metaphorical use of human concepts is transformed into a definite technical terminology. Thus, when we talk about intelligence in the context of AI research, we do not refer to the specific kind of intelligence

as exhibited in human practice, but to general intelligence, e.g. as defined in theories of ideal rationality.

But if the operationalization of concepts from the human realm leads to a determined technological terminology, it may also lead to ambiguity. Given the emergence of the terminology, it should not be surprising to find a specifically defined technological meaning of a concept besides one or more non-technological meanings under the same name. This is the case for the much-discussed concept of *autonomy*. Autonomy as a technological feature is less pertinent to AI, but an important topic of robotics; I will discuss it here as an example for a certain kind of anthropomorphism implied by terminology and resulting from the operationalization of human concepts. Additionally, the concept of autonomy also has a prominent ethical meaning, occasionally leading to a confusion of technological and ethical aspects. Fortunately, the ambiguity of autonomy is easy to clarify, but still persistent, especially in non-professional discourse.[5]

The common concept of autonomy has several roots in philosophy, political thinking, ethics and psychology. It might seem like there is a common semantic core to all the different meanings of 'autonomy'. But even if there is, it should be noted that the uses of the concept in different fields differ substantially. When we talk of autonomy as a feature of technological systems like robots or AI, we should be very careful to not mix up its different meanings. For robotics, a common definition of autonomy is the following: "[We] define 'autonomy' in robots as the capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time" (Bekey 2012, 18).[6] This definition provides a useful outline, even though the technological concept of autonomy has not yet been fully operationalized, i.e. there is no general theory of technological autonomy. For example, while there are systematic accounts of different qualitative levels of autonomy, as yet there is no discrete quantitative measure for autonomy applicable to all possible kinds of systems.[7]

One crucial aspect of the deep semantic difference between the concept of

5 For clarifications of the similarly ambiguous concept of agency (see Johnson & Noorman 2014).

6 It is worth noting that autonomous systems are not necessarily linked to technologies of Machine Learning. It could be argued that a self-learning machine represents the theoretical maximum of autonomy, but this idea is not without problems.

7 For the field of Human–Robot–Interaction (HRI), Yanco and Drury (2004) define a measure for autonomy in terms of the time a robot can function without human intervention. But, as Beer *et al.* (2014) note, this approach is lacking: "Although this idea of measuring the time of intervention and neglect is useful, it has limitations. Amount of time between human interventions may vary as a result of other factors, such as inappropriate levels of trust (i.e., misuse and disuse), social interaction, task complexity, robot capability (e.g., robot speed of movement), usability of the interface or other control method, and response time of the user" (Beer *et al.* 2014, 86). Consequently, a measure in terms of the time of intervention and neglect might not tell us anything about the autonomy of the robot in question. Additionally, according to an alternative approach to autonomy in HRI, a robot should be considered more autonomous if it is capable of more complex interactions. As a result, Beer *et al.* (2014) propose a measure that does not provide discrete levels, but only „some general comparative indication" of levels of autonomy.

autonomy of *persons* and the concept of autonomy of technical artifacts is particularly important to note: While personal autonomy marks a normative concept, technological autonomy involves no normative demands. In other words, there are *ethical* implications marking the difference between the autonomy of machines and the autonomy of persons (Christaller 2003; Müller 2014). Without addressing the theory behind this claim, its point might become clear in a thought experiment: Imagine a mobile robot that can autonomously clean the floor of a private home. The robot is able to deal with obstacles and doorsteps, it can dispose its collected dirt into the garbage can and it finally finds its way to the charger on its own. Now imagine that the robot's behavior is erratic in a very peculiar way: It never cleans a small carpet that happens to represent the colors of the national flag. For some reason, it always moves around the carpet, as if something is preventing it from moving over it and effectively cleaning it. To fix this quirk, the user sends the robot to a service engineer, who dismantles it, erases and restores its memory and finally reprograms it by changing some parts of the code. After refining its autonomous capabilities, the robot is sent back to the owner. Working on the floor, the robot is finally able to clean the disregarded carpet. The colors of the national flag shine again.

Now imagine a person doing the job of cleaning the floor, behaving similarly to the robot before being fixed. The person immaculately vacuums the whole floor, but always leaves the small carpet in a dirty state. Although the employer of the cleaner argues with him about the unattended spot, there is no change in the cleaner's behavior. Now he might have his reasons, or maybe he just prefers not to clean the carpet without providing any reasons, but the employer is in any case morally bound to accept the cleaner's behavior. The person is free to do as he pleases. The employer himself is morally not allowed to force the cleaner to change his behavior; she may try to change his mind by rationally discussing the matter with him, but not by any other means. In particular, anything like the treatment of the robot is out of the question – we would find it unacceptable to involve a neurosurgeon by asking him to 'reprogram' the cleaner's brain. This would violate the cleaner's personal autonomy.

This comparison shows that autonomy is a normative concept when it is applied to the decision-making of persons. We are morally obliged to respect the personal autonomy of others. On the other hand, there is no moral component in the technological concept of autonomy. Autonomous artificial agents are tools we use to automate certain processes; if it does not work as intended, there is no moral reason to keep the artificial agent in its unintended state.

According to this difference between the normative concept and the technological concept of autonomy, it seems reasonable to claim that no level of technological autonomy will lead to an emergence of personal (normative) autonomy, i.e. personal self-determination. The technological autonomy of a robot cannot by itself provide a reason for an approximation to the human realm. Even if technological autonomy enables automation of certain decisions, the normative impact of these decisions lies with the

persons programming and using the robot (Remmers 2020). Consequently, the ethical difference between personal and technological autonomy provides no reason to assume a difference between the moral status of (autonomous) artificial agents and any other technical artifacts. We should not approximate our treatment of artificial agents to our treatment of persons just because we use the same word 'autonomy' in reference to both entities.

But does this argument work in general? To put it in ethical terms: Is it wrong to treat robots in the same way we would treat conscious creatures in *any* case? We might recall our earlier answer to this question: If the ontological question (as presented in the first section) would be confirmed – i.e. if artificial agents actually *were* intelligent as opposed to simulating intelligent performance – this fact might provide reasons for us to treat artificial agents just like other beings with a mind. After all, a different concept of autonomy might be involved in this scenario. But if we follow our earlier argument to bypass the ontological question as purely speculative, there might still be further reasons for a special treatment of artificial agents, as outlined in the following section. And autonomy, even in its morally neutral technological meaning, is an important part of it.

## 4. Anthropomorphism in Collaboration: How Should Machines Collaboratively Interact with Humans?

Issues of anthropomorphism are currently discussed in the context of Human-Robot Interaction (HRI), but also in relation to certain scenarios of human-computer interaction (HCI) involving AI. One important goal of HRI and HCI is to enable collaborative interactions between robots or computers and humans. Collaborative interactions in HRI / HCI are interactions that are designed with a distinct similarity to common human–human interactions in mind (Chandrasekaran & Conrad 2015; Terveen 1995). Collaboration is defined as a process where two or more agents interact with each other to achieve shared goals.[8] An example in HRI is direct physical interaction of human-robot 'teams' working together in close physical proximity – something that would be extremely dangerous and harmful with traditional (non-collaborative) industrial robots.

The basic idea of collaboration in HCI and HRI is inspired by visions originating from Science Fiction. But in contrast to the idea of an 'artificial human' as known from many science fiction scenarios, we do not find many examples of a clear physical resemblance between robots and humans[9] in the context of HRI. Think of the purely functional design of industrial robots or of autonomous robotic vacuum cleaners. The popular image of the humanoid robot shaped in the image of man is rare in comparison to the many different kinds of robots in use or in development. To achieve collaborative interaction in HRI, a

8 Here, I am following the taxonomy of human–robot interaction proposed by Onnasch, Maier, and Jürgensohn (2016), who distinguish collaborative interactions from cooperative interactions and coexistence.

9 Or animals, for that matter (Johnson & Verdicchio 2018).

robot does not necessarily have to look like a human or an animal – on the contrary: One might be surprised by how little similarity is actually needed for the acceptance of a robot as a counterpart in human-robot collaboration. But even when robots do not look like humans at all, at least some kind of physical embodiment seems necessary for physical collaborative interaction.

But for collaboration in general, physical embodiment itself is not strictly necessary – more importantly, artificial agents need to exhibit a special kind of behavior that makes it easy for people to collaboratively interact with them.[10] For example, collaborative interactions presuppose some degree of perceived autonomous behavior of the artificial agent. Perceived autonomy obviously depends on the robot's technological autonomy as described in the previous section. Additionally, an artificial agent's capability to exhibit some kind of *social* behavior enables many possibilities for complex interactions with untrained persons. Consequently, social robots exemplify collaborative interaction in its purest form. These kinds of robots may not only be used for physical tasks, but also for the social or psychological effects of their behavior. In summary, physical embodiment, perceived autonomous movement and social behavior are more important for collaborative interaction than a strong resemblance of an artificial agent's shape to human shapes (Darling 2012). These features contribute to effects of anthropomorphism.[11]

An important rationale for designing collaborative interactions is the fact that non-professional users will easily find a way to interact with these systems. Collaborative interactions build on people's common knowledge when it comes to interactions. Imitation of well-known human-human interactions help to achieve this. In HRI, robots are intentionally designed for 'behavioral' anthropomorphism to support collaborative human–robot interactions.

Thinking about ethical implications of the tendency to exploit anthropomorphism in collaborative interaction, some scholars discuss aspects of deception involved in this field of application. For example, the ethical value of authenticity might be at stake (Coeckelbergh 2011). If robots actually do not think, want or feel anything – should their design or their behavior suggest that they do? Would that mean that the users are somehow deceived in an unethical way? Should users be taken to be more rational and informed? This specific issue of *authenticity* leads back to the ontological question about strong AI as discussed in the first section. The distinction of authentic and deceptive behavior of artificial agents presupposes the possibility of strong AI, or *actual* mental states as opposed to *simulated* mental states. But, as we have seen, strong AI is not a practical objective of research in AI and robotics. The distinction between real

10 In contemporary philosophy of mind, the epistemological conditions of collaborative human–robot interaction are explained in terms of the 'intentional stance' (Dennett 1987). When confronted with a robot, we ascribe intentional states like beliefs or desires to explain its behavior. This makes it easy for us to engage in (collaborative) interaction.

11 'Anthropomorphism' should be taken to include collaborative interactions with robots designed to resemble or behave like animals, because they are usually modeled after anthropomorphized animals like pets (in contrast to wild animals).

and simulated behavior is not relevant here. Therefore, it is unclear why the discussed phenomena should be taken as cases of deception. Alternatively, they can be classified as intentional and conscious engagements in performances of make-believe. As such, there are many unproblematic examples of spontaneous anthropomorphism to be found in ordinary behavior towards animals and children's dolls or in games of make-believe. As long as these tendencies stay within reason, they constitute no ground for the assumption of any serious ethical issues.

Still, as Kate Darling notes in reference to findings of her colleague Sherry Turkle, "there is a difference between the type of projection that people have traditionally engaged in with objects, such as small children comforting their dolls, and the psychology of engagement that comes from interacting with social robots, which create an effective illusion of mutual relating" (Darling 2012, 7). In other words, in collaborative scenarios of HRI, anthropomorphizing robots is not something that humans freely choose and control at will, but something that is strongly suggested and demanded by the technical artifacts themselves. It is increasingly required for the technology to work as intended. This tendency might lead to several ethical issues: emotional attachment to technical artifacts, substitution of human relationships by human-robot-relationships, new opportunities to manipulate users in a way that is not in their best interest or, last but not least, obscured responsibilities due to diffusion or overconfidence. And as it is often the case in the advancement of technology, these issues might evolve unnoticed until the use of the artifacts is entrenched to the point of no return.

Additionally, in this specific context, the tendency of anthropomorphism may eventually lead to blurring boundaries between technology and humans, or between minds and machines. Even if we deny a specific moral status of artificial agents because they do not have a mind, the emerging collaborative *relations* between humans and artificial agents may constitute reasons for acknowledging a moral status of the whole interaction.[12] This point is most obvious in the context of social interactions. If I start to treat an artificial agent as a companion who provides company, the relation of companionship is real, even if the artificial agent cannot really mind.

In this case, ethical implications are hard to dismiss, leading to new questions in philosophy of technology, anthropology and social philosophy. To address them, interdisciplinary endeavors are called for, involving robotics, ethics, philosophy, psychology and social science. It is at this point that philosophy and ethics are responsible to actively engage and participate in technological design – taking a decidedly different route than the purely theoretical engagement with riddles introduced by science fiction scenarios and futurist thinking.

12 Along the line of this reasoning, some scholars argue for a post-human revolution in ethics, demanding moral emancipation of artificial agents (Gunkel 2012; Coeckelbergh 2014). Their position is the extreme antithesis to the allegedly 'rational' dismissal of any ethical relevance in collaborative human–robot interaction.

## References


Beer J. M., Fisk A. D., & Rogers W. A. 2014. "Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction." *Journal of Human-Robot Interaction* 3(2):74-99.

Bekey G. A. 2012. "Current Trends in Robotics: Technology and Ethics." In: *Robot Ethics. The Ethical and Social Implications of Robotics* (pp. 17-34). Cambridge, Mass. – London: MIT Press.

Chandrasekaran B. & Conrad J. M. 2015. "Human-Robot Collaboration: A Survey." *SoutheastCon*:1-8.

Christaller T. (Ed.) 2003. *Autonome Maschinen*. Wiesbaden: Westdeutscher Verlag.

Coeckelbergh M. 2011. "Artificial Companions: Empathy and Vulnerability Mirroring in Human-Robot Relations." *Studies in Ethics, Law, and Technology* 4(3):1-17.

Coeckelbergh M. 2014. "The Moral Standing of Machines: Towards a Relational and Non-Cartesian Moral Hermeneutics." *Philosophy & Technology* 27(1):61-77.

Darling K. 2016. "Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior towards Robotic Objects." Calo, Froomkin, Kerr (Eds.), *We Robot Conference 2012, University of Miami*. Edward Elgar.

Dennett D. C. 1987. *The Intentional Stance*. Cambridge, Mass.: MIT Press.

Dretske F. 1994. "If You Can't Make One, You Don't Know How It Works." *Midwest Studies in Philosophy* 19(1):468-82.

Gunkel D. J. 2012. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. Cambridge, Mass.: MIT Press.

Johnson D. G. & Verdicchio M. 2018. "Why Robots Should Not Be Treated Like Animals." *Ethics and Information Technology* 20(4):291-301.

Johnson D. G. & Noorman M. 2014. "Artefactual Agency and Artefactual Moral Agency." In: P. Kroes & P.-P. Verbeek (Eds.), *The Moral Status of Technical Artefacts* (pp. 143-158). Dordrecht: Springer.

Lanier J. 2010. *You Are Not a Gadget: A Manifesto* (1st ed). New York, NY: Knopf.

Miller K. D. 2015. "Will You Ever Be Able to Upload Your Brain?." *New York Times*, Oct. 10, 2015 (retrieved from https://nyti.ms/1VLghZ4).

Müller M. F. 2014. "Von vermenschlichten Maschinen und maschinisierten Menschen." In: S. Brändli, R. Harasgama, R. Schister, & A. Tamò (Eds.), *Mensch und Maschine—Symbiose oder Parasitismus?* (pp. 125-142). Bern: Stämpfli.

Onnasch L., Maier X., & Jürgensohn T. 2016. „Mensch-Roboter-Interaktion—Eine Taxonomie für alle Anwendungsfälle." *baua: Fokus, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin*.

Remmers P. 2020. “Would Moral Machines close the responsibility gap? Reflections on Autonomous Artificial Agents.” In: B. Beck & M. Kühler (Eds.), *Technology, Anthropology, and Dimensions of Responsibility*. Stuttgart: J. B. Metzler (in press).

Russell S. J. & Norvig P. 2016. *Artificial Intelligence: A Modern Approach* (3rd edition). London: Pearson Education.

Searle J. R. 1980. “Minds, Brains, and Program.” *Behavioral and Brain Sciences* 3(3):417-24.

Terveen L. G. 1995. “Overview of Human-Computer Collaboration.” *Knowledge-Based Systems* 8(2-3):67-81.

Turing A. 2004. “Computing Machinery and Intelligence.” In: B. J. Copeland (Ed.), *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life, Plus the Secrets of Enigma*. Oxford: Oxford University Press.

Weizenbaum J. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco: Freeman.

Yanco H. A. & Drury J. 2004. “Classifying Human-Robot Interaction: An Updated Taxonomy.” *2004 IEEE International Conference on Systems, Man and Cybernetics* (IEEE Cat. No. 04CH37583) 3:2841–2846.

Peter Remmers (Berlin)

# The Ethical Significance of Human Likeness in Robotics and AI

**Abstract:** A defining goal of research in AI and robotics is to build technical artefacts as substitutes, assistants or enhancements of human action and decision-making. But both in reflection on these technologies and in interaction with the respective technical artefacts, we sometimes encounter certain kinds of human likenesses. To clarify their significance, three aspects are highlighted. First, I will broadly investigate some relations between humans and artificial agents by recalling certain points from the debates on Strong AI, on Turing's Test, on the concept of autonomy and on anthropomorphism in human-machine interaction. Second, I will argue for the claim that there are no serious ethical issues involved in the theoretical aspects of technological human likeness. Third, I will suggest that although human likeness may not be ethically significant on the philosophical and conceptual levels, strategies to use anthropomorphism in the technological design of human-machine collaborations are ethically significant, because artificial agents are specifically designed to be treated in ways we usually treat humans.

**Keywords:** AI; robotics; human likeness; anthropomorphism; ethical implications; Strong AI; Turing's Test; autonomy.

Ethics in Progress (ISSN 2084-9257). Vol. 10 (2019). No. 2, Art. #6, pp. 52-67.

DOI:10.14746/eip.2019.2.6