AI & Soc (2009) 24:181–189
DOI 10.1007/s00146-009-0208-3

OPEN FORUM

Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents

Mark Coeckelbergh

Received: 4 June 2008 / Accepted: 18 August 2008 / Published online: 6 May 2009

1 Introduction

Contemporary technology creates a proliferation of non-human artificial entities such as robots and intelligent information systems. Sometimes they are called 'artificial agents'. But are they agents at all? And if so, should they be considered as *moral* agents and be held morally responsible? They do things to us in various ways, and what happens can be and has to be discussed in terms of right and wrong, good or bad. But does that make them agents or moral agents? And who is responsible for the consequences of their actions? The designer? The user? The robot? Standard moral theory has difficulties in coping with these questions for several reasons. First, it generally understands agency and responsibility as individual and undistributed. I will not further discuss this issue here. Second, it is tailored to human agency and human responsibility, excluding non-humans. It makes a strong distinction between (humans as) subjects and objects, between humans and animals, between ends (aim, goal) and means (instrument), and sometimes between the moral and the empirical sphere. Moral agency is seen as an exclusive feature of (some) humans. But if non-humans (natural and artificial) have such an influence on the way we lead our lives, it is undesirable and unhelpful to exclude them from moral discourse. In this paper, I explore how we can include artificial agents in our moral discourse, without giving up the 'folk' intuition that humans are somehow special with regard to morality, that there is a special relation between humanity and morality—whatever that means. Giving up this view happens if we lower the threshold for moral agency (which I take Foridi and Sanders to do), or if we call artefacts 'moral' in virtue of what they do (which I take Verbeek to do in his interpretation of Latour and others) or in virtue of the value we ascribe to them (which I take Magnani to do). I propose an alternative route, which replaces the question about how 'moral' non-human agents *really* are by the question about the moral significance of *appearance.* Instead of asking about what kind of 'mind' or brain states non-humans really have to count as moral agents (approach 1), about what they really do to us (approach 2), or about what value they really have (approach 3), I propose to redirect our attention to the various ways in which non-humans, and in particular robots, *appear* to us as agents, and how they influence us *in virtue of this appearance*. Thus, I leave the question regarding the moral status of non-humans open and make room for a study of the moral significance of how humans perceive artificial non-humans such as robots and are influenced by that perception in their interaction with these entities and in their beliefs about these entities. In particular, I argue that humans are justified in ascribing *virtual* moral agency and moral responsibility to those non-humans that *appear* similar to themselves—and to the *extent* that they appear so—and in acting according to this belief.

Thinking about non-humans implies that we reconsider our views about humans. My project in that domain is to shift at least some of our philosophical attention in moral anthropology from what we really are (as opposed to non-humans) to anthropomorphology: the human form, what we appear to be, and how other beings appear to us given (our projections and recreations of) the human form. I want to make plausible that it is not their intentional state, but their *performance* that counts morally, and that we can gain

M. Coeckelbergh (✉)
Department of Philosophy, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: m.coeckelbergh@utwente.nl

by moving from a discussion about artificial intelligence to a discussion about artificial performance.

2 The trouble with the standard account of moral agency and responsibility

Let me start from what I call the standard account of moral agency and moral responsibility, which is ascribed on the basis of moral agency.

Agency refers to the ability to act. Usually action is understood as more than 'doing'; it depends on, and is the result of, a mental state such as 'desire', 'intention', 'decision', etc. Agency need not involve freedom or rationality. The mental state and the action can be 'not free' but caused by an external entity; the decision need not be rational, etc. On this account, a cat can be an agent, but a plant cannot. Following Himma (2007), on whose useful summary of the discussion on agency I draw in this section, we can make a further distinction between natural and artificial agents. If it has intentional states that lead to (some would say: cause) 'performances', a 'highly sophisticated computer could be considered an agent' (Himma 2007, p. 5), and some advanced robots could also be considered an agent if they meet the same condition. Thus, on this account of action and agency there has to be (1) an intentional mental state and (2) this state has to 'cause' a performance.

A further step, then, is *moral* agency. Ideas about moral agency depend on the view taken on what counts as 'moral' or 'morality'. Some say that the action has to be governed by moral standards, that we have moral rights and duties. Most of us agree that moral agency entails that the agent can be held accountable for its action, that is, it can be held responsible. Let me summarize what I take to be the standard account of moral agency and moral responsibility.

Moral agency depends on at least two conditions and one common precondition (Himma 2007, pp. 14–18). First, one has to have the capacity to freely choose one's acts; the agent's behaviour is not compelled by something external to. Furthermore, this requires that the person deliberate, or has at least the capacity to do so. Free choice also presupposes that the agent be rational. A second condition is that one has to know the difference between right and wrong. This requirement is often understood as knowing how to apply moral concepts and principles. A common necessary (but not sufficient) precondition for these two conditions is that the agent has the capacity for consciousness and self-consciousness. Punishment requires that it is possible 'to produce an unpleasant mental state' (Himma 2007, p. 17), and free choice supposes the capability of conscious deliberation.

There are various theories of *moral responsibility*, which rely on similar conditions: a moral-libertarian condition and a moral-epistemological condition. Most refer back to the two Aristotelian conditions as put forward in the *Nicomachean Ethics*: (1) the agent has to act freely and (2) needs to know what (s)he is doing. Sometimes the first condition is called autonomy or self-control: I can change and control my internal states; they are not put into me by an evil demon or a mad scientist (to use examples from the literature). Animals are supposed not to meet these conditions, and therefore they are not to be blamed or punished. They do not 'deserve' it (of course, in concrete life they often are blamed and punished—especially if they *appear* to meet the two conditions. I will return to this observation). But what about artificial agents? Could highly 'developed' or 'evolved' artificial agents be said to act freely, understand the requirements of morality, and know what they are doing? Are (very) young children moral and responsible agents in this sense? Are criminals who are diagnosed 'psychopathic' moral agents? It is not difficult to see that this account of moral agency and moral responsibility meets many challenges in the face of practical problems. Not only is it hard to attribute freedom and consciousness to natural or artificial agents *in principle*, the particularly hard question is the practical one of determining whether or not any *particular* natural or artificial agent is free and conscious. As Himma notes, 'philosophers have yet to solve even the problem of justifying the belief that other human beings than ourselves are conscious ('the problem of other minds')' (Himma 2007, p. 23), so how can we tackle the problem of other *artificial* 'minds'? More generally, my question is: how can we talk about the morality of non-humans within this framework, given that the attribution of a moral status depends on conditions which are difficult to prove and (in many instances) unlikely to be met?

Note that the standard account is, within its own boundaries, highly incomplete in its description of what morality consists in: morality does not only require the application of principles, but also the use of moral *imagination*, which allows us to explore the potential consequences of our actions, create novel action options, improvise, and put ourselves in the place of the other (Johnson 1993; Fesmire 2003; Coeckelbergh 2007). But this addition to the theory does not solve the problem with regard to the moral status of non-human agents, since imagination is a mental operation. The problem of (other) minds persists.

The problem even gets worse when we consider the crucial role of technology and artefacts in shaping our contemporary society and existence. For instance, Bruno Latour's actor-network theory and his thinking on modernity (Latour 1993) are reflections of the importance of materiality in our lives. Technology and artefacts 'do'

things to us—sometimes good things, sometimes bad things—and co-shape how we live our lives (Verbeek 2005). But does that render them moral? According to the standard theory, they are not even agents, let alone moral agents. There is no evidence that computers have mental states (some philosophers even continue to doubt that *humans* have such states—this is the so called problem of 'other minds'). Some predict that computers could get mental states in the future, or that the internet will become conscious. But so far, these technologies are 'mindless'; yet, they exert much influence over our lives. A speed bump is not an agent, since it does not have mental states. Yet it directs our behaviour. With some imagination, one could say that someone else had the intention to slow the traffic down, and *delegated* this intention to the artefact, the speed bump.[1] But this does not imply that the speed bump itself has an intention—or any mental state for that matter. And it does not help to apply the term 'delegated' responsibility, since there are always unexpected outcomes possible, so in our practice of responsibility ascription we cannot refer to the designer alone. The problem of responsibility ascription remains: who or what is responsible?

Even if artefacts *could in principle* have mental states, consciousness etcetera and therefore would be considered as (moral) agents, then we still have no way to find out whether or not a *particular* artefact has these properties, and this gets the standard theory in trouble since it requires such proof. On the standard account, without such proof we cannot ascribe responsibility to, say, an artificially intelligent 'mass murder' robot; the robot and other entities of its kind remain outside the moral sphere altogether.

But why holding on to the standard theory? Some try to pull artefacts into the moral domain by giving them the label 'moral' in virtue of their ability to cause good or bad things or in virtue of their being the object of moral attribution. Let me explain these two options. The first is to recognise that artefacts, things have moral consequences, some of which are 'delegated' to them by the designer, some of which are unintended. In *What Things Do* Verbeek calls this 'the morality of artefacts' (Verbeek 2005).[2] At first sight, there seems no good objection to applying the label 'moral' to artefacts in this specific sense. However, as I argued, the fact that things *do* something to us, does not itself warrant calling them *moral agents*, or holding them *responsible*. We do not want to say that a speed bump is 'responsible' for slowing us down. If one chooses to talk about the 'morality' of things, as Verbeek (2005) does, then this 'solution' is at the cost of running against the intuition that humans and morality have a special, exclusive relation. Moreover, Verbeek's approach applies to all artefacts, but does not account for the difference we experience between what a speed bump does to us and what a humanoid robot with (some) artificial intelligence does to us. The first is about the consequences artefacts have for our behaviour and our lives. But we do not experience these artefacts themselves as being 'moral'. The second—so we experience or imagine—is about what something that is 'more than a thing' does to us. It appears to us as an agent. How can we better conceptually grasp this intuition?

A different way of applying the label 'moral' to things is not to focus on the transfer of intention and responsibility from subjects (humans) to objects (artefacts), but on the transfer of value. We humans not only design artefacts, but also give value to objects, artefacts, and other non-humans. In this sense we also make them 'moral'. I take Magnani (2007) to follow this route in his recent book. However, this is not the sense of 'moral' we imply when we talk about 'moral agents': we want to say something about its (apparent) agency, not about what is being done to it. Moreover, I cannot see how the mere observation that an object is valued—is given value—could be a reason at all to ascribe responsibility to it.

In sum, with regard to the problem of responsibility, these two alternative approaches do not really solve the problem. And apart from the already mentioned difficulties, an underlying major problem persists: we need to know 'what really is the case' in order to ascribe responsibility. While we no longer have to discuss about the real mental states of animals, robots, and objects, we still have to find out that is the case about the real intentions of the designer or the real value people assign to things. This renders the ascription of responsibility difficult, since we need access to the 'mental states' of people. And this was exactly the main problem I detected in the standard approach.

The solution I propose is to shift our attention from the real to appearance. As long as we define the problem of moral agency and responsibility of non-human, artificial agents as a problem of mind ('other minds'), then it remains a mystery how even *humans* are able to act morally and responsible, and we will never be able to conceptually grasp what *already* goes on between humans and non-humans with regard to morality. Verbeek's turn to what things *do* remedies this problem, but his conclusion that we therefore should talk about 'moral' artefacts or the 'morality of things' goes at the cost of diluting the anthropocentric meaning we like to give to the terms 'moral' and 'morality'. And Magnani rightly turns to the

[1] For a discussion of Latour's concept of *delegation* see also Verbeek (2005).

[2] The term was first introduced by Hans Achterhuis, who thought we should 'moralise' things, that is, delegate our moral intentions to them.

subject—we give value to things—but this does not take seriously the fact that things *do* morally relevant things in the sense that they have morally relevant consequences. Moreover, we not only attribute value to things; in practice, but also attribute agency and responsibility to non-human entities. This is the relevant problem I started from. Finally, part of the difficulty emerges from an approach to (moral) philosophy that first aims to 'get the concepts right' and then wants to apply it to (moral) practice. I propose to start from experience and practice. In particular, I wish to take seriously the anthropocentric practice of agency-attribution and responsibility-attribution to non-humans. We already experience and treat some animals and perhaps some advanced robots as moral agents. How can we justify and make sense of this experience and practice without taking recourse to the standard approach? What kind of practices and experiences do I have in mind, and what concepts do we need to understand and evaluate them?

3 Appearance, experience, and practice: towards a conception of virtual agency and responsibility

Let us pause for a moment and consider how we experience and treat other *humans*. The standard account of moral agency and moral responsibility starts from a kind of Cartesian 'mind' problem: we might doubt that our own thoughts, beliefs, desires, etc., are really ours. Is there a demon which deceives me, and pulls the strings? I might falsely believe that I act freely. This doubt is then projected onto 'other minds': do they also act 'freely'? Do they have minds? Are they conscious at all? Can we, therefore, be sure that they are moral agents, or agents at all? Perhaps your wife is a robot, your colleague a zombie, and your neighbour a projection in your mind? Perhaps we must doubt all external reality, as Descartes did? Perhaps we live in 'The Matrix'? In real life, however, people seldom contemplate such an issue, and if they did so frequently, they would rightly run the risk of being considered mad by their fellows. Instead, people go on to interact with each other, presuming that the other is a free, moral agent who can and should take moral responsibility. Of course, sometimes we distrust others with regard to their moral status and virtue. For example, the jurisdiction in some murder cases hinges on the question whether or not the person can be considered responsible for his (most of the time it is a man) deeds. But if we ask the question and try to find out, we—professionals and others—do this, not on the basis of an investigation of their mind ('where can we find free will?'), but rather on the basis of how they *appear* to us and what we *experience* when we interact with them. Furthermore, if we wonder what other people think, we use our capacity for empathy: we put ourselves in their place, we imagine how it would be to stand in their shoes. We do not really penetrate into the 'depth' of their minds, and there is no need to do so in moral practice. We interact with others, treat them, ascribe responsibility to them, and blame them on the basis of how they *appear* to us, not on the basis of what kind of mental states that person really has—if we could every know that at all.

In our relation with *non-humans*, then, we use a similar *as if* approach, which turns out to be sufficient to support the (quasi)moral and social dimension of our dealings with them. We treat pets such as dogs and cats as if they have their own 'will' and 'thoughts'. We (will) interact with humanoid robots as if they are human. We need not know their 'mental states' for blaming them, for treating them as companions, or even for loving them. Both the ascription of agency and of responsibility are, in practice, independent of the real. I coin the terms *virtual agency* and *virtual responsibility* to refer to the responsibility humans ascribe to each other and to (some) non-humans on the basis of how the other is experienced and appears to them. This concept accounts of our present and future talking about some non-humans (including artificial agents) in moral terms, and sustains our moral practices.

Consider again how standard moral theory tries to tackle the problem of responsibility of artificial agents. Himma writes: 'Free will poses tremendous philosophical difficulties that would have to be worked out before the technology can be worked out; if we do not know what free will *is*, we are not going to be able to model it technologically' (Himma 2007, p. 22). I believe it is exactly the other way around. Free will has to be 'technologically modelled' first, in robot/AI design or in imagination, before we can fully work out the philosophical difficulties. First we have to design or create in fiction artificial agents. We can then observe or imagine that some of them will *appear* to have a free will and other mental features and mental states, that they have virtual moral agency. We notice how humans interact with them on this basis, and acknowledge this in our moral theory. For instance, if some people would really start to love robots, then this will force us to discuss about the concept of love. Similarly, if some people started to treat some humanoid robots as if they are moral agents, then we would need to think about our concept of moral agency.

But does this mean that virtual agency and responsibility are descriptive concepts only? I do not think so. *Real* responsibility, as I shall call the concept of responsibility put forward by the standard account, allowed us to discriminate between good and bad ways to ascribe moral responsibility, or rather, it was *meant* to do so, but it did not very well succeed given its dependence on insight in mental states. *Virtual* responsibility can also discriminate between good and bad ways to ascribe moral responsibility,

but instead of making that evaluation dependent on (searching for) the existence of 'real' mental states, it relies on the *appearance* and *experience* of mental states, and on (observing or imagining) the practices based on that appearance and experience. Thus, instead of asking whether or not a criminal really had the capacity of free will at the moment of his deed, we must acknowledge that we cannot know the answer to that question. Rather, we can only go by appearance: we can try to imaginatively construct the crime scene and the history of interaction between criminal and victim, put ourselves in the mind of criminal and victim, etcetera. And there are good and bad ways of doing that. Some perceptions and imaginative (re)constructions are more adequate than others. This is what happens in practice, of course. Legal epistemology pretends to go for reality, but it can only reach appearance. Nevertheless, it manages to reach conclusions on responsibility and blame that are more or less accepted by society. Legal and non-legal responsibility ascription in society relies on 'outer' performance, not on 'inner' reality. As for our interaction with non-humans, we observe that in interaction with pets, for example, humans project themselves into their animals, anthropomorphise them. But this need not be a moral problem. It is perfectly acceptable—morally and socially—to proceed in this way, if and when there is indeed a *similarity in form*, in appearance between the two. We evaluate the animal in terms of its performance: it may or may not appear to have free will. The aesthetic-phenomenological dimension of agency and responsibility ascription already has considerable and sufficient normative power to sustain the moral life—whatever the 'reality inside' may be.

Of course, virtual responsibility, based on virtual agency, should be followed (if at all) by *virtual* blame and punishment, not real blame and punishment. That is, if we ascribe virtual responsibility to a human, a dog or a robot, then we should not attempt to *really* blame or punish them. Instead, it suffices that there is the *appearance* of what Himma calls an 'unpleasant mental state'. Thus, we could create humanoid robots that appear to be unhappy when we blame them or 'punish' them. And if people wish to 'blame' their dog and act in such a way that the dog appears to be unhappy, why not? *To the extent that* they appear as (quasi-)social beings, a speech act may suffice for that purpose, and, more important, such punishment must meet the condition that they *appear* as responsible, moral agents. To me and most people both pets and current robots appear differently: they do *not* appear to us as virtual moral agents, which means that the condition for virtual responsibility (and therefore blame and punishment) is not met. In the future this might change. But note that for designers the requirement is considerably lower than that of the standard account: the appearance of moral agency suffices, the robot need not have real mental states, a real 'mind'—if that were possible at all. Humans, by contrast, *can* meet the condition in many instances and cases, though not always. And the appearance of punishment, which knows many varieties, should suffice. Punishment should not be 'symbolical', since this would mean we suppose a link with the real, of which we cannot be sure, but rather *performant*, i.e., able to create the appearance of suffering. This, at least, is in tune with our moral practices (and sometimes the appearance of regret is held more important than the appearance of suffering). But in any case, in court and elsewhere we have to go by appearance, we can never know for sure what the person really thinks or feels at the time of the deed and at the time of punishment.

One may object: why should we punish at all, if responsibility is virtual? And does the above justification for the 'theatre of punishment' not amount to giving in to lust for blood? Several justifications have been given for punishment of humans, and many of them are at least morally dubious (Honderich 1969). But what about virtual moral agents and virtual punishment? Consider the two main theories: retributivism and consequentialism (including utilitarianism). Retributivism says that we should punish someone because (s)he deserves it and in proportion to the severity of the crime. With regard to humans, it is hard to find a good justification for this view, but it can be *explained* as a satisfaction of revenge, or, as Honderich (1969) has argued, a satisfaction of grievance. Now in case of virtual desert and virtual punishment, the 'advantage' is that such feelings of revenge or grievance on the part of humans—and 'the lust for blood'—can be satisfied without necessarily doing harm to anyone. However, this argument assumes that satisfaction of such feelings is acceptable as a moral good and a social aim. This brings us to the question what kind of society we want. This is the focus of consequentialist arguments: punishment is justified if it contributes to a better, more humane society, e.g., by preventing crime or by providing other gains for society. Again, as an argument for punishment this is weak with regard to humans (for lack of empirical evidence), and that does not change if we consider virtual punishment of virtual moral agents. Our main task seems to make sure that such entities do not harm humans and society but rather benefit them. Unless the satisfaction of feelings of revenge and grief counts as a benefit, virtual punishment does not seem to contribute to that aim. Instead, it appears to contribute to the de-humanisation of society.

Apart from the problem of justification, there is the further question if robots or other intelligent entities *can* be punished at all. The notion of virtual punishment solves the problem that artefacts cannot feel pain or suffer, since this is not required for virtual punishment. Since I shifted the focus from the capacities of the virtual agent to perception

by *humans*, the requirement is different: I expect that *if* there is the appearance of moral agency, there is also a capacity to be virtually punished. But again, I think this is unlikely, and certainly at the moment.

But if there are little or no artificial agents and robots that appear to us as virtual moral agents, how relevant is this discussion for today? First, the artificial agents of the future are designed now, step-by-step. If we care for a pro-active ethics that intervenes with its evaluation at the design stage rather than when the artefact is used, we better think about the ethical problems now. Second, this discussion serves to clarify our own self-understanding as moral beings. In that respect, one of the further lessons to learn is that if responsibility ascription happens on the basis of appearance, we should be at least very careful when and before we morally evaluate *humans*—inside and outside the court. Given our reliance of appearance and our lack of strong arguments for punishment, we should be very cautious and careful in our evaluation of whether or not the person is responsible, and whether or not punishment is the best way to respond to what happened.

For thinking about artificial agents such as robots, this turn towards appearance does not release us from difficult issues concerning responsibility. For instance, is a military robot responsible for shooting humans? How is the responsibility distributed between the robot and certain humans? My proposal is not to abandon such a reflection, but rather ask a slightly different question. Instead of asking whether or not the robot is conscious, rational, free-willed, etc., let us turn our attention to how the robot appears: does it 'exhibit' such capacities as supposed in humans? If so, then regardless of whether the robot really has these capacities and mental states, we should ascribe moral agency and moral responsibility to the robot. Then we can ask further questions, such as: does it appear to share this agency and responsibility with other artificial or natural agents, how is the precise distribution of these apparent properties?

Now, let me return to the conditions for moral agency to further discuss the plausibility and relevance of artificial moral agency. Let us grant that it is likely that now and in the near future some robots will have virtual *agency*; that is, they will appear to us as acting on the basis of a mental state. But virtual *moral* agency, it seems, is a different matter. One may take the view that if a robot appeared to have a mental state at all, this state would not *seem* to be caused free-willed. By extension, it would not seem that the robot is applying 'the moral law', and/or using its moral imagination. Indeed, it seems unlikely that such a virtually free, imaginative robot will ever be built, or will ever be built in the near future. This implies that, unless and until anyone builds an entity that appears to have moral agency, we should blame the users and the designers for morally bad consequences. If something goes wrong, they must face the challenge of showing that they (the humans) are *not* responsible, since—in contrast to robots—as humans they will 'prima facie', at first sight, *appear* to us as morally responsible agents.

In the theatre, a good actor is one who really *appears* to have the desires, beliefs, emotions, etc., of his character. We do not ask about the reality of the desires, beliefs, etc. But in 'real' life we usually do not ask such questions either. In 'non-fiction' life, we interact with humans and non-humans and ascribe responsibility to them without investigating the reality of the other's mental states. If this is true, we must adapt our moral frameworks in so far as they are based on an unrealistic epistemological assumption and demand, and at least supplement them with a moral-aesthetic criterion. If our (inter)actorship has a virtual side to it, then we must also recognise the virtual dimension of moral responsibility, blame, and punishment.

The analogy between the theatre and social life has been employed before, although usually it has not been extended to the moral life. In *As You Like it*, Shakespeare wrote his famous words 'All the world's a stage; and all the men and women merely players'. And in sociology social role is an important concept, and Latour has introduced the notion of script in science and technology studies. These concepts deserve further discussion; however, here I will make some associations of my own in order to further explore the analogy in order to support my argument concerning virtual moral agency.

As they enter our world, some artefacts appear as new players on the scene. The play changes, and this forces us to review our own role. Perhaps we first have to improvise. We can also try to give a certain role to them to write their script. We can try to (re-)direct the play. In any case, what happens is based on how the human and non-human actors appear to one another. The masks matter socially and morally. They blame one another, they praise one another. Together they write a narrative. They act it out. They are actors and co-directors, they follow scripts that are written (be)for(e) them, and they change the script and improvise.

Note that applying the notion of play does not imply that the moral or social life is not serious. There are winners and losers. Consider how we treat non-human animals today: some we treat like princes, others like raw materials. How we treat them, which role we assign to them, depends on their masks and appearance. Towards already existing artificial actors and indeed humans, many of us act in the same way: cute robots are loved, service robots are used and ignored. My point is not that we *should* (not) act in that way, but that we must recognise how important the virtual dimension is in social life. If we do so, then we can take that into account when we re-write the play, look critical at traditional scripts, and introduce acts and scenes that are

morally superior to the current ones. We can try to see the same actors in a new light. We can try to empathise, imagine how it is to wear the mask of the human or non-human other. And we can change stage and settings—institutional and others—in order to promote some actions rather than others.

Importantly, new *artificial* actors do not enter the stage from nowhere. We design them. We can give them the mask, the appearance we want. Since we love ourselves as humans, it is very likely that we will give them our own mask. Or the mask of the animals that remind us of our own, human infants.

Note that an additional advantage of my approach is that it can not only account for virtual agency in the so-called 'real' world, but also in the virtual worlds created by ICT. In this context, the standard theory is entirely helpless, since no-one would seriously contend that a robot avatar (an avatar that is not connected to a real human being but run by software) has a 'mind' of its own, has 'mental states', etc. Still, we are happy to ascribe virtual responsibility in some instances when immersed in a computer-generated virtual world, without that this experience depends on mental states on the part of the robot avatar. AI adepts may think it will be possible to create mental states in the future, but in any case the theory of virtual moral agency and virtual moral responsibility can cope with both situations. What counts in practice is that the designers of such robot avatars try to create the illusion that we interact with a real human agent. This suffices for generating and sustaining the experience of virtual moral agency and the corresponding practice of virtual responsibility ascription. Again, we can apply the normative criterion: if they manage to create the illusion of virtual moral agency, then we are justified to hold these 'entities'—which are virtual by design—also virtually responsible for what they do to us.

4 Mind-less morality versus virtual mind morality (an ethics of appearance)

Let me now further distinguish my approach from other responses to the limits of the standard theory of moral agency and moral responsibility. I already discussed two alternatives that focus on the 'morality' of artefacts. But what about artificial agents? One could object to my approach: Why not ascribe *real* responsibility to some of these agents in some cases? Are there reasons to do so? Verbeek and Magnani do not really account for the intuition that the moral status of such agents is not exhausted by a description of what they do (Verbeek) or by the value we give to them (Magnani). They appear to us as *more* than a thing, an object. Can we give a 'higher' moral status and perhaps moral responsibility to such non-humans? Why should we take humans as the standard, the model for moral agency anyway? A response that still needs further discussion, then, is what I called in my introduction 'lowering the threshold of moral agency'. This I take Floridi and Sanders to do. I will use this section to distinguish my own approach from their view in order to further clarify my own account.

In their influential paper 'On the Morality of Artificial Agents', Floridi and Sanders (2004) share my view that we should move away from the 'traditional approach' which requires us to find out whether or not agents have mental states, feelings, emotions, etc. For them, however, the major problem lies in the traditional *anthropocentric* conception of agency. Instead of focussing on the human mind, therefore, they propose a 'mind-less morality' (Floridi and Sanders 2004) which allows us to analyse whether or not a system is an agent in terms of its interactivity (response to a stimulus by change of the state of the system), autonomy (the ability to change without such a stimulus), and adaptability (the ability to change the transition rules by which the state is changed). On this basis, they argue that artificial agents can be accountable sources of moral action, without being responsible or exhibiting free will (Floridi and Sanders 2004, p. 351). In other words, they separate moral agency and moral responsibility, and conceive of the former in (what they see as) non-anthropocentric terms. In their view, dogs and artificial systems can be moral agents, without being morally responsible for their actions (Floridi and Sanders 2004, p. 368). They claim that artificial agents can be 'morally accountable as sources of good and evil—at the 'cost' of expanding the definition of morally charged agent' (Floridi and Sanders 2004, p. 372). They conclude that their account, by analysing entities in non-anthropocentric terms, manages to 'progress past the immediate and dogmatic answer' to the problem (Floridi and Sanders 2004, p. 375).

Let me clarify my own approach by objecting to their assumptions that anthropocentrism is necessarily dogmatic and something that must be overcome, and that their approach succeeds in doing that.[3]

First, it is not clear to me why their systemic approach to agency is supposed to be entirely non-anthropocentric. As far as I know, the systems metaphor is a human-made term for (doing something to) human-made things. We design systems and apply the label 'systems' to artefacts and combinations of artefacts. We decide that it is a combination, or we make the combination. In order for there to be

[3] A comprehensive discussion of Floridi and Sanders would include scrutinising their distinction between moral accountability and moral responsibility and other elements of their view. I limit my discussion to those aspects of their view and their approach that help me to clarify my own view and approach.

a 'system', there has to be a (human) subject that orders, assembles, applies a method to, and makes distinctions and relations between objects. Technology itself is deeply 'anthropocentric' and intrinsically bound up with human culture. There is no artificial or cultural system separate from the practices that construct, imagine, and live it.

Second, I believe there is a deeper disagreement between us about the relation between philosophical understanding and common sense understanding, and about the method of philosophy. Their approach is part of philosophy as (meta-) science, which is a philosophy of suspicion: *you* 'ordinary' people live in the illusion that humans, dogs, and (future) artificial agents have a moral status, whereas *we* scientists–philosophers show and know that these are all systems which should be analysed in terms of levels of abstraction (LoA) and other technical terms that really show us the moral similarities and differences between these entities. In taking (or at least suggesting) this approach, they are in agreement with the mind-morality philosophers they criticise. The approach I suggest in my work on robotics and artificial agents (AA), by contrast, starts from observed or imagined human-AA interaction, aiming at taking seriously how humans experience and co-shape such an interaction, including their potential ascription of human-like agency and human-like responsibility. Starting from this platform, I ask the normative question if, how, and when this interaction and ascription could be justified. The dogmatism *I* want to progress beyond, then, lies in the assumption—made by both 'mindful' and mindless morality philosophers—that we can have full access to reality, regardless of whether this reality is described in terms of mental states or system states. My suggestion is that we can permit ourselves to remain agnostic about what *really* goes on 'in' there, and focus on the 'outer', the interaction, and in particular on how this interaction is co-shaped and co-constituted by how AAs appear to us, humans.

Part of the 'folk' way (to use a term of the science-oriented philosophers) to think about morality is that if we perceive moral agency in the other, we also hold that other responsible. It is also part of our 'folk' moral psychology that if we one day are 'deceived' into believing that the artificial 'system' we meet is human, we will treat and interact with that entity as if it was human. We will consider 'it' to be a moral agent and hold 'it' responsible. The famous Turing-test (Turing 1950) and the Chinese Room thought-experiment (Searl 1980) are used in discussions about what really is the case about artificial intelligence and mind (can AI be conscious? Can it have morality? As if morality is something that can be built-in). But such thought-experiments depend on appearance: what counts is how the entity appears to us in our interaction with that entity, not on what is 'really' in the room (which stands for the mind or the system).

Thus, the imaginative projection of anthropomorphic features onto beings of which we know that they are not human, such as pets and pet robots is not necessarily problematic, as long as we ascribe moral status and moral responsibility *in proportion to* the apparent features. Following such an 'ethics of appearance' is what we do in practice: we use moral-aesthetic measures in our interaction with human and non-human others, not moral-epistemic measures. Not (only) the 'content' but the form, per*form*ance of the other matters to us in concrete (quasi-) social interaction. Let us turn our attention from epistemology and philosophy of mind, directed to objects and minds (trying to find out the truth about artificial agents), to the moral-social question, directed to relations between social others (trying to find out how we can live with them). My approach is also expansive, but in contrast to Floridi and Sanders (and perhaps in contrast to Verbeek), it does not expand the definitions of moral agency we employ. Instead, I propose to stretch the conceptual boundaries of the social sphere: let us expand the definition of 'others' from humans to (humans and) non-humans. When we interact with others, we have only appearance to go by. Using our senses and our imaginative projection and empathy, we may well live in 'illusion' as far as agency and responsibility is concerned. But it is not so much the 'truth' about which entities 'really have' agency and responsibility that matters in the moral life understood as the social life. Rather, it is the *appearance* of the other that matters with regard to our experience and understanding of the other's moral status and responsibility, and with regard to the practices based on that experience and understanding.

Perhaps we should be content that such a 'shortcut' via appearance is available to us from an evolutionary point of view: if we first had to 'get the concepts right' and 'get the facts right', our moral, interactive and (quasi)social life would freeze to death. This would mean that human life would come to and end. And to many of us, pathetic anthropocentrics as we are, leaving the world to our artificial offspring is an unthinkable and bleak prospect—whether or not these orphaned entities would count as moral and responsible agents.

5 Conclusion: the perception and design of virtual agency and virtual responsibility as artificial performance

My discussion about the limitations of the standard account of moral agency and moral responsibility resulted in a sketch of an alternative approach that shifts the emphasis from ('inner') reality and truth towards ('outer') appearance and performance. Now if in (quasi)social interaction

with humans and non-humans appearance indeed plays such an important moral role, then the model of agency and responsibility ascription must be changed from the model of science to the model of art, in particular what can be called an art of perception, imagination, and design of performance. Instead of trying to find out the truth about the 'content' of the other's mind, it suffices for our moral practices of moral agency and moral responsibility ascription that we develop our capability to perceive, experience, and imagine the form and performance of the other. In concrete practice and experience, we must try to find out if this form and performance of the other (and ourselves) is congruent with the model of human moral agency we aspire to, a moral-aesthetical ideal that continues to haunt much of the moral philosophy and philosophy of action literature: the ideal of a free-willed, self-conscious, rational agent, which we should supplement with at least the capacities for moral sensitivity and moral imagination. Some humans (at times) and all current artificial agents are likely to fail this test of moral appearance, but if and when they meet the criterion, now or in the future, we are justified in holding them virtually morally responsible for what they *do* to us in virtue of their appearance (and therefore for what they seem to *intend* to do to us).

With regard to the ascription of moral agency and moral responsibility in interaction with artificial agents (AAs), I draw the following conclusion. For the 'spectator', deciding about ascribing responsibility to such agents is not about getting the facts about the (content of the quasi-) other's mind right; instead, it is about experiencing, sensing and perceiving a form and performance which we so far only experienced in interaction with humans. For the designer, the challenge is to create an artificial 'actor' that produces this appearance. What counts with regard to moral status and moral responsibility, then, is not so much the AA's artificial intelligence (AI), but its *artificial performance* (AP) with regard to the appearance of agency. This way of putting the problem is closer to what many designers really aim at in practice. It is also a concept that does justice to the quasi-social aspect of interaction between humans and AAs. It invites us to draw AAs into the sphere of moral consideration by taking them seriously as quasi-social entities who are *already* part of our sphere of social consideration in virtue of their appearance. Philosophical reflection on AAs and morality, therefore, should be practiced not only as a philosophy of mind, but, at least also, as a social philosophy and a moral aesthetics. To the extent that such entities appear to us as moral agents (if they ever do at all), we should generously welcome them in the conceptual dwellings we built for us, humans. If they (will) have a role in our social plays, and to the extent that they (will) have that role, we must consider them as our fellow actors, and assign virtual agency and virtual responsibility to them without hesitation.

References

Aristotle Nicomachean Ethics, William DR (trans.). Clarendon Press, Oxford, 1908

Coeckelbergh M (2007) Imagination and principles. Palgrave Macmillan, Basingstoke

Fesmire S (2003) John Dewey and moral imagination. Indiana University Press, Bloomington

Floridi L, Sanders JW (2004) On the morality of artificial agents. Minds Mach 14:349–379

Himma KE (2007) Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Available via Social Science Research Network. http://ssrn.com/abstract=983503

Honderich T (1969) Punishment: the supposed justifications. Hutchinson, London

Johnson M (1993) Moral imagination: implications of cognitive science for ethics. The University of Chicago Press, Chicago

Kant I (1785) The moral law: Kant's groundwork of the metaphysic of morals, Paton HJ (trans.) Routledge, New York, 1991

Latour B (1993) We have never been modern. Porter C (trans.). Harvard University Press, Cambridge

Magnani L (2007) Morality in a technological world: knowledge as duty. Cambridge University Press, Cambridge

Searl J (1980) Minds, brains and programs. Behav Brain Sci 3(3):417–457

Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460

Verbeek PP (2005) What things do—philosophical reflections on technology, agency, and design. Penn State University Press, Penn State