Laukyte, M. (2017). Artificial agents among us: Should we recognize them as agents proper? *Ethics and Information Technology*, 19 (1), pp. 1-17.

DOI: 10.1007/s10676-016-9411-3

Artificial agents among us: Should we recognize them as agents proper?

Migle Laukyte[1]

migle.laukyte@uc3m.es

[1] UC3M Conex – Marie Curie Fellow, Departamento de Derecho Privado, Universidad Carlos III de Madrid, 15.02.73, Calle Madrid 126, Getafe, 28903 Madrid, Spain

Abstract

In this paper, I discuss whether in a society where the use of artificial agents is pervasive, these agents should be recognized as having rights like those we accord to group agents. This kind of recognition I understand to be at once social and legal, and I argue that in order for an artificial agent to be so recognized, it will need to meet the same basic conditions in light of which group agents are granted such recognition. I then explore the implications of granting recognition in this manner. The thesis I will be defending is that artificial agents that do meet the condi-tions of agency in light of which we ascribe rights to group agents should thereby be recognized as having similar rights. The reason for bringing group agents into the picture is that, like artificial agents, they are not self-evidently agents of the sort to which we would naturally ascribe rights, or at least that is what the historical record suggests if we look, for example, at what it took for corporations to gain legal status in the law as group agents entitled to rights and, consequently, as entities subject to responsibilities. This is an example of agency ascribed to a nonhuman agent, and just as a group agent can be described as non-human, so can an artificial agent. Therefore, if these two kinds of nonhuman agents can be shown to be sufficiently similar in relevant ways, the agency ascribed to one can also be ascribed to the other—this despite the fact that neither is human, a major impediment when it comes to recognizing an entity as an agent proper, and hence as a bearer of rights.

Keywords

Agency · Group agent · Artificial agent · Rights · Responsibility · Personhood · Rationality

Introduction

In this paper, I discuss whether in a society where the use of artificial agents is pervasive, these agents should be recognized as having rights like those we accord to group agents. This kind of recognition I understand to be at once social and legal, and I argue that in order for an artificial agent to be so recognized, it will need to meet the same basic conditions in light of which group agents are granted such recognition. I then explore the implications of grant-ing recognition in this manner.

The thesis I will be defending is that artificial agents that do meet the conditions of agency in light of which we ascribe rights to group agents should thereby be recognized as hav-ing similar rights. The reason for bringing group agents into the picture is that, like artificial agents, they are not self-evidently agents of the sort to which we would naturally ascribe rights, such as human beings for instance. A group agent is an example of agency ascribed to a nonhuman agent, and just as a group agent can be described as nonhuman, so can an artificial agent. Therefore, if these two kinds of non-human agents can be shown to be sufficiently similar in relevant ways, the agency ascribed to one can also be ascribed to the other—this despite the fact that neither is human, a major impediment when it comes to recognizing an entity as an agent proper, and hence as a bearer of rights.

As just mentioned, the argument for recognizing artifi-cial agents as having rights will depend on their meeting the conditions of agency by virtue of which group agents are themselves so recognized. We will therefore have to spell out a conception of agency stating what those

conditions of agency are. In so doing I will be taking a two-pronged strategy: On the one hand, I will block out a conception of agency on which an agent is any entity that is both *rational* and *interactive*, meaning that it has a capacity to reason rationally (in a way shortly to be defined) and can interact intelligently with other agents (in a way that will also be explained); on the other hand, I will show that from this conception of agency as competence we can derive two attributes of agency—namely, responsibility and personhood—that we can ascribe to agents capable of engaging rationally and interactively with other agents.

So what I ultimately want to say is that if an artificial agent can be described as (*i*) rational and (*ii*) interactive, then we can ascribe (*iii*) responsibility and (*iv*) personhood to it, and consequently we can recognize it as having rights based on those capacities and attributes: four conditions of agency satisfying which an artificial agent can be grouped among other agents having similar rights. And since in so recognizing artificial agents as agents proper we have to understand them as acting in an environment shaped by social and legal rules, I will finally be discussing what this may entail for the way our relation to them ought to be framed.

But before we start, I would like to briefly introduce the idea of rights as they will be used in this paper. I will not enter here into the discussion of different conceptions and theories of rights: except to clarify two premises from which I will be proceeding. The first one is that I am taking the view that whatever rights we should think it reasonable to ascribe to artificial agents, these rights will have to be specifically tailored to the features and abilities these agents are going to have (on what I will be describing as the competence approach). And the second premise will be a working definition of rights as "entitlements which is incumbent upon others to acknowledge and respect" (Jones 1994, 1). This notion of rights picks up Hohfeld's (1917) idea that a right is made up of claims, privileges, powers and immunities[1] and that with such rights come specific sets of correlative duties and responsibilities. Thus, for example, if we grant an agent a right to form a contract, we should also encumber that agent with a duty to perform the same contract.

Having briefly shed the light on the idea of rights that I argue artificial agents could be subject to, I will now move to the first condition of agency, namely rational agency and explore how artificial agents satisfy it.

First condition of agency: rational agency

I premise this discussion by noting that a rational agent is one that can *act* rationally. That is, an agent is an entity that *acts* in an environment, so the kind of rationality we are interested in is the kind we can ascribe to it by looking at what it *does* in that environment.

In order to introduce a concept of rational agency, I will rely on what List and Pettit (2011, 19ff.) call a "basic account of agency," on which an agent is anything that can (a) sense the way things *are* in its environment (through its representational states), (b) know the way that environment *should* be (through its motivational states), and (c) proceed to act in the same environment so as to fill that gap between what *is* the case and what *ought to be* the case. On this basic account, then, an agent has three sorts of features: (a) an ability to represent the environment such as it is (through its representational states); (b) an idea of how it wants the environment to be (through its motivational states); and (c) an ability to process those states so as to bring about the required changes. List and Pettit (ibid., 20) summarize that threefold idea by saying that something is an agent if it "has representational states, motivational states, and a capacity to process them and to act on that basis."

It follows from the foregoing definition of rational agency that if something has representational states and motivational states but processes them in a way that cannot make rational sense to us, then that thing cannot count as an agent. For example, if a thermostat has a representational state corresponding to a temperature of 15 °C (current temperature) and a motivational state corresponding to 25 °C (set temperature), but processes those two states in such a way as to set the temperature at 10 °C, then we cannot take it to be a rational agent (assuming the thermostat is not broken and that it is designed to bring the temperature to the level at which it is set). The same can be said to apply to an artificial agent: As long as something can have representations and motivations through which it acts in its environment, and it can act so as to realize those motivations in light of its representations, then that thing counts as an agent in this basic sense.[2]

In order to make sense of this idea that an agent can be viewed as rational and intentional even without ascribing a rational thought process to it or a set of intentions, we need

[1] In each of these specifications, a right gives one a normative ability to do or not do something: This can be the ability to demand something from someone (rights as claims), or the freedom to do something that is not prohibited (rights as privileges), or the ability to modify a legal situation (rights as powers) or not be subject to the powers of others (rights as immunities). For a discussion, see Hohfeld 1917 and Jones 1994.

[2] A caveat before we proceed is that the thermostat example just introduced should not be taken to mean that an artificial device is rational just because it correctly executes the instructions it is designed to execute. Nor should the motivational states we attribute to it be taken to mean that it somehow "wants" or "intends" to do what it does. The example is rather intended to illustrate that we can explain an agent's actions as if it were rational and intentional, without saying that it is a rational agent driven by actual intentions.

to bring in Daniel Dennett and point out that we are looking at agents not from the physical stance of the natural sciences but from what he has famously called the intentional stance, the "strategy of interpreting the behavior of an entity (person, animal, artifact, whatever) by treating it *as if* it were a rational agent who governed its 'choice' of 'action' by a 'consideration' of its 'beliefs' and 'desires'" (Dennett 2009, 339). The key to that view is the "as if" phrase which is used to underscore that we don't have to ascribe mental states to agents in order to describe them as rational or as having beliefs and desires: We just have to be able to analyze their behavior as *consistent with* the way a rational agent would act, to which end we need not con-sider their mental or physical states, or rather we can do so as long as it is clear that these states are introduced as heuristic devices, without having to look at what actually goes on in the "minds" or "brains" of such agents.

In the same vein, List and Pettit (2011, 29) note that rationality as previously defined need not be a component of agency so long as we can interpret the agent as having acted consistently with some piece of reasoning, and since we are *interpreting* what the agent is doing, we need not concern ourselves with identifying any mental or physical state or series of states corresponding to such reasoning. We just need a set of criteria in light of which to judge whether a given course of action is rational, and to this end List and Pettit (2011, 24) introduce three standards of rationality necessary for an entity to qualify as an agent. These are an ability of the agent to connect its attitudes (its representational and motivational states) with (a) its environment, (b) one another, and (c) the actions by which the agent intervenes in its environment.

With the previously mentioned caveat that an agent's attitudes do not necessarily amount to mental states, List and Pettit provide a compelling case that group agents can be said to act rationally when they satisfy these standards of rationality, and an argument can be made that the same holds for artificial agents. This is because whatever an artificial agent does, its action can be judged as either consistent or inconsistent with a piece of reasoning (or with the standards in light of which any course of action counts as rational), and to this end we need not concern ourselves with the agent's "inner workings," that is, with the way in which an agent arrived at the course of action that we are interested in judging as rational or otherwise. We can see, then, that even though group and artificial agents both lack a conscious mind in the way a human agent can be said to have one, their action can nonetheless be judged as rational or otherwise, *as if* a mind were behind the reasoning that led to the given outcome in question. We regularly judge group agents in light of standards of rationality, and there is no reason why we cannot also hold artificial agents to the same standards.

Having said that, rationality alone does not account for the whole of agency: Even if an agent satisfies a set of accompanying standards of rationality, we cannot yet consider it an agent for the purpose of attributing rights to it or holding it accountable, as we do with group agents. A thermostat may pass a test of rationality, but we certainly wouldn't confer rights on it for that reason alone, nor would we hold it responsible for failing to comply with such a test. The reason why rationality cannot alone define an agent worthy of moral consideration (an agent recognized as having rights) is that agents typically do not act in a vacuum but rather *interact* with other agents: Their action unfolds in an environment shaped by the action of other agents, and the resulting interaction thus turns out to be essential to agency itself. We should therefore consider the second component of agency, consisting in an agent's ability to interact sensibly with other agents.

Second condition of agency: interactive agency

It was just suggested a moment ago that it is unrealistic to think of an agent as a lone entity acting in an empty space: I am therefore going to posit that an agent (*i*) needs to be rational and (*ii*) must necessarily be interactive (no matter how simple its agency). A rational agent is one whose action is consistent with some piece of reasoning understood to make some logical sense (even if the agent is not *itself* reasoning in the sense of making inferences and suchlike); an interactive agent is one that, by virtue of its action, necessarily interacts with other agents (however minimal or constrained such interaction may be). I should note that while the first condition is a *desideratum* of agency (in that we *want* agents to be rational), the second condition is a *matter of fact* of agency, in that an agent is ipso facto interactive in virtue of its acting in any environment (for any environment is going to be inhabited by other agents that either act on the agent or stand affected by its action).

This second condition of agency (its interactivity) can be illustrated by looking at the way List and Pettit (2008, 75) characterize group agents , defining them as groups of networked individuals who (a) understand themselves as part of a group and (b) act in respect of that group in such a way that the group can be recognized as rational in much the same way that an individual can be so recognized. It may seem obvious that the individuals making up a group must necessarily somehow interact if they are to be recognized as forming a group, but the point here is that such interaction is inescapable, and in turn the group must inevitably interact with other (individual or group) agents the moment it does anything as a group.

Now we can turn our attention back to artificial agents and point out three main features they share with group

agents: (1) Both can be understood as rational, or as satisfying some criteria of rationality; (2) both are nonhuman; and (3) both have some kind of social ability, meaning that they can interact with one another or with human beings. This in turn means that artificial agents engage in a variety of activities, and what all these activities have in common is that they "conceptually presuppose the existence of other agents and various social institutions" (Tuomela 1984, 1).

The reason for focusing on what artificial agents have in common with group agents is that we are already accustomed to seeing the latter as having rights, and if we can show a strong enough similarity between these two kinds of agents, we have a reason to recognize the first kind (artificial agents) as having rights like the ones we ascribe to the latter kind (group agents).[3] It is in particular the second common feature that makes this for an interesting comparison: Like artificial agents, group agents are non-human (they are not persons per se), and this has histori-cally made it difficult to recognize them as having rights like the ones we ascribe to human agents. So if we can show that artificial agents share with group agents a set of features in virtue of which the latter are recognized as agents proper, then we should see that it is inconsistent to recognize such rights for one kind of agent (group agents) while denying them to another (artificial agents). And the fact that both are nonhuman can then be seen to fade into the background as irrelevant to whether they should be owed such recognition.

The two significant features of agency so far discussed that artificial agents share with group agents is that both are rational and both are interactive. So what is it about interactivity that can make it a feature of agency significant enough to warrant the conclusion that interactive agents can be recognized as having rights?

I answer this question by making two observations. The first is that (*i*) Whenever any set of agents interact on the basis of some "code" they execute or some piece of reasoning they act on, the action of some agents is going to affect that of others; (*ii*) when this mutual effect is significant enough, it is going to be either harmful or beneficial to some of the agents involved; (*iii*) whenever any harm or good is involved in any interaction between agents, the question of right and wrong comes up; and (*iv*) whenever the question of right and wrong comes up, we can ask whether someone is rightfully entitled to the good they benefited from or is responsible for the harm they suffered. Of course the agent needs to be autonomous in some way in order for these questions to be asked sensibly (and I address that question later on in "Structural difference" section), but for the time being it is enough that we recognize how these questions can come up and how they relate to that of rights.

The second observation (which I will be developing at the end of "Fourth condition of agency: personhood" section) is that when agents interact, they are likely to do so by playing different *roles*, and when roles are involved we can ask what is expected of the agents that fill them and what those agents need in order to fill those roles properly. This, too, is essentially a question of rights, and it is in virtue of the roles ascribable to interactive agents that we can begin to bring that question into view.

Of course, even when the interaction casts agents in different roles and affects them in ways that are either good or bad, we still do not have all the conditions of agency needed to recognize them as having rights. To this end we need to introduce the third and fourth conditions of agency, namely, responsibility itself and personhood. This is what we will do in "Third condition of agency: responsibility" and "Fourth condition of agency: personhood" sections, showing how the responsibility and personhood that List and Pettit ascribe to group agents can also be ascribed to artificial agents.

Third condition of agency: responsibility

We can now look at List and Pettit's account of responsibility so as to see how it applies to both group and artificial agents. Responsibility is described by them in a straightforward way as a concept dependent on an underlying notion of good and bad behaviour (which we are assumed to have an intuitive grasp of): "If what was done is something bad, then the agent is a candidate for blame; if it is something good, then the agent is a candidate for approval and praise" (List and Pettit 2011, 154).[4]

List and Pettit go on to specify three conditions for an agent to be fit to be held responsible (ibid., 158), and in so doing they complement the idea of good and bad with that of right and wrong. We can see this in the first condition, that of normative significance:

(*i*) Normative significance simply means that the agent is facing a normative or moral situation, that is, a situation involving a normatively significant choice or option, or "the possibility of doing something good or bad, right or wrong" (ibid.).

[3] I should note that the parallel between group agents and artificial agents is not new (see Solum 1992; Singer 2013). List and Pettit (2011) and Pettit (2007) seem to reject that parallel, since they consider the agency of a "bare-bones" artificial agent (a very stripped-down robotic device) in contrast to the full agency of group agents. But as can be appreciated from the way artificial agents were just defined, I understand them to comprise a class much more inclusive than that of robots.

[4] This is a standard position on moral responsibility: See Himma (2009).

The second condition is a capacity for normative judgment, requiring an agent to have an *understanding* of the situation just described:

(*ii*) Agents can be said to have a capacity for normative judgment if (a) they can single out the features of a situation that make it moral or normative, and (b) they understand that what they do in light of a situation so framed carries moral or normative consequences; that is, the agent in question understands that different ways of handling a situation have different outcomes (the agent may have a concept of harm, for example) and that not all of these outcomes carry the same weight (it may have a concept of moral desert or fair distribution, for example), and it will therefore not treat those outcomes equally, or at least it will treat them in such a way that we can infer an understanding of their moral significance on the agent's part. For example, the agent understands that there is a normative problem involved when different people make a claim on the same resource and that different distributions of those resources lead to morally or normatively different outcomes. In this example, an agent can be said to have a grasp of all three of the normative concepts mentioned parenthetically: harm, moral desert, and fair distribution. In other words, the agent recognizes that harm can be done to someone by not giving them the resources they claim; it can recognize that they can claim those resources only if they meet criteria such as need (this would be a way of modelling moral desert); and it can distribute those resources accordingly (fair distribution). This means that even if we use mental or intentional concepts to describe the behaviour of agents that lack any mental or intentional states properly so called, we can analyze their behaviour as consistent with that of agents that do have such states and that use them as a basis on which to make normative judgments.[5]

(*iii*) The third condition is the control requirement, meaning that the agent must be able to exercise control over the options available: There is no moral responsibility involved in the face of a situation we can do nothing about. This is a fairly intuitive idea and seems plausible at face value, but it does run into some difficulties when it comes to spelling out exactly what it means to "control" your options, at least for a human agent: It may well be that we are indeed in control of a situation when the choice is presented to us in the context of a moral problem to be solved in the abstract—are you going to "sacrifice one person's life in order to save several other lives"?—but then we may no longer feel in control if we are the person who is actually doing the sacrificing, because our emotions will get in the way and our moral judgment will change accordingly (Greene et al. 2009, 364). And, generally, control is not an all-or-nothing affair, as if every morally significant situation we are faced with is one whose outcomes we either control or do not control. So this third condition needs to be taken with a grain of salt, and each case will accordingly have to be judged on its own merits.

The three conditions of moral responsibility can be summarized in the statement that you can't be held responsible for some state of affairs unless (*i*) that state of affairs is the outcome of a choice that can bring harm or loss to yourself or to other people, (*ii*) you understand what the implications of that choice are, and (*iii*) you actually had an opportunity to make that choice—but a couple of more points need to be mentioned in that regard before we proceed.

The first point is that, as noted, List and Pettit (2011) rely on an underlying notion of what counts as good and bad or as right and wrong behaviour, and it was mentioned parenthetically that we are assumed to have an intuitive grasp of those two notions, that is, we slip them in as unchallenged premises. This is actually a gaping hole in List and Pettit's account of responsibility. But I suspect that the reason why we are asked to make those assumptions is that the two concepts at hand—the right and the good—are so fundamental to moral philosophy and have been so widely discussed over the course of history that any satisfactory account of them would take us on a long detour from which it would be difficult to come back, and even if we did firm up a thoroughly reasoned out theory of the good and the right, chances are that when it comes down to the nitty–gritty of practical judgment in specific cases (everything from broad policy decisions to what we should have for lunch), different people (or agents) reasoning from the same theory will arrive at different conclusions about what ought to be done. So it's much simpler to assume that we already know what's good and what's right in any given instance, without having to justify those judgments. And even though this is certainly a shortcut, it doesn't mean that we *cannot* justify the judgments we make: We can, and we probably also *should* do so whenever the issue at hand is not so simple as how to divide a pie so that we each get our fair share, that is, whenever disagreement can arise about what is morally good or right.[6]

[5] I should point out, as previously suggested, that while a capacity for normative judgment is an essential condition subject to which responsibility can be ascribed to an agent, we also have to look at the *roles* agents play in the environment in which they interact, for this is essential in figuring out the *kinds* of responsibilities that can be ascribed to them and the *consequences* that should follow as a result of the agent failing to fulfil those responsibilities. The question of roles is discussed at the end of "Fourth condition of agency: personhood" section.

[6] For other criticisms about the fitness to be held responsible, see Tuomela (2011).

The second point is that, as much as these three all-encompassing conditions may seem broad and abstract, they are reflected in the standards that lawyers and judges use to resolve the very practical cases that arise in tort liability. Consider the conditions that must be satisfied in a suit in order to prove that someone was negligent (and so is to be held responsible for some state of affairs). The person bringing the suit (the plaintiff) "must show four things: (a) there was a duty imposed on the defendant in favor of the plaintiff, (b) the defendant breached (violated) that duty, (c) the duty was the proximate (natural and foreseeable) cause of the harm, and (d) plaintiff suffered damages" (Emerson and Hardwicke 1997, 376). Take out any one of List and Pettit's three conditions of responsibility and you can no longer account for the legal elements of negligence. So, if we try to do away with List and Pettit's first condition (normative significance), we end up looking at a state of affairs that was going to happen anyway (there was no choice involved and so no choice can be pointed out as the source of the harm caused): This means that we cannot even begin to entertain the legal idea of (a) a duty (which presupposes a course of action that must be taken when others are possible, and so a choice) or the idea of (d) damages (which explicitly means that harm was done to somebody, the plaintiff). Likewise, if we try to do away with List and Pettit's second condition (a capacity for normative judgments), then we cannot make sense of the legal idea of (b) a breach of duty, for this idea presupposes an ability to understand what is at stake when we act in such a way as to violate an expectation (which in turn presupposes a choice). And, finally, if we try to do away with the third condition (the control requirement), we end up looking at a situation that was not caused through any agency, and so we cannot make sense of the legal idea of (c) proximate cause, which presupposes that someone can do something (i.e., can control the situation) in such a way as to bring about the harm in question.[7] This is not to say that the doctrines the law has evolved are thereby justified simply because they are the law, but it does suggest that List and Pettit's conditions of responsibility model the presuppositions of responsibility, that is, the factors that need to be taken into account before we can even begin to ascribe responsibility to an agent.

List and Pettit explain how all three conditions can be met by group agents. We will not be entering into this explanation here, but I should point out that implicit in these conditions of moral responsibility is the assumption that if an agent is to be held morally responsible, it must be in a position where it can make choices, for that is what it means to act in a morally significant situation. So, built into this account of an agent's responsibility is the standard view that this concept is closely bound up with that of freedom: An agent can be said to be responsible only to the extent to which it is free, "such that no matter what you do, you will fully deserve blame should the action be bad, and fully deserve praise should the action be good" (Pettit 2001, 12). This standard view of moral accountability is essentially the aforementioned Kantian principle that *ought* implies *can*, coupled with the corollary that *can* implies freedom. I mention this because, on the one hand, the principle is central to the argument I am making about what it means for an agent to be held responsible, but at the same time it is not always clear how the freedom required of a responsible agent is to be specified: As an abstract concept it means that an agent has options in dealing with the situation at hand, but as a practical matter these options may not be easy for the agent to see or choose. So there is often much interpretation that goes into deciding whether an agent can be held responsible on this Kantian principle. The difficulty involved in applying the principle, however, does not warrant the conclusion that we should invalidate it.

The argument I will now be making is that the aforementioned requirements for responsibility can also be satisfied by *artificial* agents, or that there is, in principle, no condition of responsibility that group agents can satisfy but artificial agents cannot.

The first condition (normative significance) simply requires an agent to find itself in a situation to deal with which a choice needs to be made that somehow carries moral import, and there is no conceivable situation in which this might apply to a group agent but not to an artificial one. An example might be a military robot operating in a warzone and chancing upon a child soldier who is preparing to shoot at civilians: Any choice the robot will make in such a situation—and in particular the decision whether or not to fire at or otherwise disable the child soldier so as to save civilian lives—will have moral ramifications, regardless of whether any moral considerations factor into the robot's decision-making. Of course, this is not a morally straightforward choice, because there are valid reasons on both sides of the argument, but it is nonetheless a morally significant choice, and one the artificial agent cannot escape.

Next we can turn to the second condition for responsibility, the capacity of an agent to form a normative judgment. This is actually a twofold condition in List and Pettit's description, for in the first place an agent must have access to the relevant facts or evidence on which a moral choice hinges: This is a general problem not distinctive to any specific kind of agent (group or otherwise). It is more

[7] Another example where List and Pettit's three conditions of responsibility find a counterpart in the law is in the legal concept of force majeure, which excuses a party from responsibility for nonperformance ascribable to events beyond that party's control.

an epistemological problem than an agential one. So we can assume that the agent has access to all relevant information, and at this point we can focus on the second part of this condition, which is that an agent must be able to recognize information as morally relevant and be able to draw moral conclusions from it: as List and Pettit (2011, 158) put it, an agent must be able "to form judgments on normative propositions" expressed in terms of "it is right that X." In the example of the child soldier, a robot agent must have a capacity to understand what is morally at stake in the situation, deciding which is the lesser of the two evils, namely, using deadly force on a child or allowing civilians to be targets of the child.

Clearly, this can be a difficult predicament for any human, let alone for a group or an artificial agent. List and Pettit argue that this ability can be ascribed to group agents through the individuals who form the group, but the same argument cannot be applied to artificial agents, because in this case there are no constituent individual agents through which such an ability can be exercised.[8] As a practical matter, however, we should consider that if robots can learn from experience by observing human behaviour, and if the beha-viour they are learning from is consistent with standards of rightness, then these robots can likewise be said to act in a normatively correct way. And even though they may not be able to understand what is right about the behaviour they are mimicking, there is no inherent feature of their internal language or programming that should preclude an ability to so reason. So there is much room for improvement. And, in addition, some authors claim that future artificial agents can be morally superior to us because these agents "would lack an evolutionary past like ours that dooms us to a core of bad behaviors" (Dietrich 2011, 531).

With that said we can turn to the third condition: Having encountered a normatively significant situation (first condition), and having appreciated its moral import and consequently formed a practical judgment about how it ought to be solved (second condition), an agent is required to have the *control* needed to *act on* that judgment. With group agents, the problem is how the group can be said to be in control of an action when it is the *individuals* in the group who make all the decisions on the group's behalf and materially carry out those decisions. The problem, List and Pettit point out (2011, 161), is parallel to a classic problem in the philosophy of mind, that of multi-level causality, where the question arises: Is it at the neuronal level that action is controlled or is it at the mental level, namely, the level of an individual's intentional attitudes? And their argument is that, just as the neurons cannot be said to rob the individual of causal control over his or her actions, so the individuals who make up a group cannot be said to rob the group of control over its decisions about how to act. A similar problem arises in regard to artificial agents, for it can be asked whether control over their actions rests with designers or with the design itself.[9] And here, too, an argument can be made that just as our neurons do not rob us of the control we can exercise over our actions, so an artificial agent's designers cannot be said to deprive the agent of all control over its actions: The agent will still maintain some autonomy (a subject I expand on in "Two kinds of autonomy" section), and so will continue to exercise some control, for otherwise it wouldn't be recognized as an agent to begin with.

So, in summary, the three conditions necessary for an agent to be held morally responsible can be argued to apply to group and artificial agents alike, in that artificial agents can (a) find themselves facing a normatively charged situation, one whose outcomes are morally significant;(b) judge the situation in ways that take those normative features into account; and (c) exercise a degree of control in making decisions on that basis. The condition that is most difficult for an artificial agent to meet is clearly (b), since artificial agents do not engage in the practical rea-soning needed to work through the implications of a nor-matively significant situation, but for one thing there is no reason to think that they can't develop such an ability in the future, and for another—on the intentional stance previ-ously introduced in "First condition of agency: rational agency" section—their behaviour can be interpreted as consistent with practical reasoning, and may even be pre-dicted by attributing practical reasoning to them, without having to invoke something like a thinking mind behind that reasoning.

We have considered the grounds on which responsibility can be ascribed to artificial agents, so let us take up the question of their personhood, again drawing a parallel between artificial and group agents.

[8] This structural difference will be taken up in "Structural difference" section.

[9] I should note here that this parallel between neurons and individuals, on the one hand, and individuals and groups, on the other, is itself up for debate. It would be rejected on an incompatibilist view such as hard determinism or metaphysical libertarianism. The former would argue that there is no free will in virtue of which an individual or group agent might control its actions—for that control is only mechanistic (Illes 2005, 45)—such that the question of responsibility wouldn't arise in the first place. The latter, for its part, would grant that responsibility is an issue, but only for human beings and only if they have "a freedom to originate action uncaused by prior events and influences" (ibid.).

Fourth condition of agency: personhood

Just as we backtracked to a basic concept of agency and a broad account of responsibility in making the argument that there are grounds on which artificial agents can be held responsible for their actions, so a similar strategy can be adopted in considering their personhood. To this end I ask the simple (albeit philosophically fraught) question, What is a person? The question is relevant because, depending on how it is answered, we can extract consequences about the way we ought to relate to artificial agents as a society, and so how we should deal with them in the law. Let us begin by noting, in this regard, that a person, according to List and Pettit (2011, 173), is essentially what I would call a social-relational being, someone with a capacity to function in a social setting governed by a system of mutual expectations. As they put it, a person can "be party to a system of accepted conventions, such as a system of law, under which one contracts obligations to others and […] derives entitlements from the reciprocal obligations of others. In particular, it is to be a knowledgeable and competent party to such a system of obligations."

This sociality of personhood is an important point I will be developing in what follows, but before we get there we should clear two methodological errors out of the way as we approach the question of how personhood might be ascribed to an agent. The first error is essentially an application of the naturalistic fallacy, and the idea is that we cannot ascribe personhood to an agent just because the agent somehow has the makings of a natural person.[10] The second error consists in taking what List and Pettit (2011, 170–71) call the intrinsicist view, which as the name suggests would have us ascribe personhood on the basis of what an agent intrinsically is, by determining the "essence" of that agent. The problem with the naturalistic fallacy is that an agent may resemble a human being and yet have none of the features on which basis we would call a human being a person (a case in point might be a mannequin). The problem with the intrincisist view, on the other hand, is that the question of essence is too speculative: It can easily yield abstract principles subject to any number of interpretations and unlikely to be of any practical use.

Once we clear those two errors out of the way, we can focus on a third approach in deciding whether an agent has personhood. This approach, very much in line with the discussion so far, is based on what List and Pettit call the performative conception of personhood, the idea being that in deciding whether or not the entity before us is an agent, we should consider not its likeness to a natural person (the naturalistic fallacy) or what that agent essentially *is* (the intrincisist view) but what it does or *can do* within a range of possibilities.[11] If you'll recall the definition of an agent introduced at the outset on the basic account of agency, an agent was someone who can act in the world so as to bring about the desired changes, with an emphasis on what an agent can do, precisely the emphasis that distinguishes the performative account of personhood. Now, next to that emphasis we can bring in a new one by focusing not only on the ability to act in the world with a view to changing it (in accordance with an agent's motivational states) but also on the ability to act in a *social* world. On a performative approach, then, an agent can be considered a person if it can act in *both* of the worlds just mentioned: the physical world of actual possibility and the social world of interaction, a world framed by rules, principles, and conventions about what may and may not be done and by what is required or expected of one (List and Pettit 2011, 174). In both respects, an agent qua person is like a human or group agent, not in the sense that there is a natural resemblance between the two, but in the sense that these agents behave in ways that enable them to be part of a system of mutual obligations and accepted conventions: This is something that nonperson agents cannot do, since they do not have an awareness of what is expected of them or of what they can expect from others. So the thing that makes an agent a person is this capacity to operate within a system of obligations and conventions, and for this reason, as List and Pettit argue (ibid.), whatever the agent in question is, if it can engage with humans and group agents within a commonly established framework of conventions, then it can be regarded as a person.

With that view of personhood in place—call it the performative-relational view—we can now consider

[10] Although it is a fallacy to proceed on a basis of likeness to human beings in ascribing personhood to an agent, there is no denying that humans do react differently in their interaction with a robot when the robot *looks* human. As the roboticist Daniel Wilson observes (Singer 2009, 405), we unconsciously make judgments based on a robot's form and "care differently about a humanoid robot versus a dog robot versus a robot that doesn't look like anything alive."

[11] Yet another approach to personhood is the interest-based one offered by Briggs (2012), who takes List and Pettit's view of personhood to mean that "a person is the sort of thing to which it is appropriate to assign conventional rights" (Briggs 2012, 289) and thus suggests that we look to *interests* as the basis on which to assign those rights, the idea being that it makes no sense to ascribe rights to something (say, a rock) if that thing "cannot benefit from those rights" and so cannot be said to have an interest in them. This idea that something ought to have rights to the extent that it can benefit from them calls up the competence approach (because implicit in that idea is that of an underlying capacity, or ability, to benefit from the rights in question), but at the same time, an interest-based approach would be more restrictive in its ascription of rights than would the interrelational approach I will be introducing shortly, for if we take interests as a basis of ascription, we may not be able to contemplate the idea of the environment, for example, as having any interest in protection and so as a subject of rights.

whether it can be extended not only to group agents (as List and Pettit do) but also to artificial agents. For if we can do that, we will be justified in conceiving artificial agents as persons in that performative-relational sense. The argument for that extension can be made by combining two observations. The first one is that what the performative-relational view essentially proposes is a variety of the Turing test—for it tests for an intangible quality *x* by seeing whether the entity in question can do *y*—and the second one is that the same basic idea underpins a class of approaches (I would accordingly call them functional or Turing test approaches) that have been used to test for qualities that are either akin to personhood or identical with it.[12] So the argument, in one long breath, contains three premises as follows: If (a) the performative-relational view belongs with a broader class of approaches that all rely on the same insight to test for an intangible quality *x*, (b) the quality being tested for is either personhood itself or something closely associated with it—like consciousness or intelligence, such that an entity recognized as having one quality or combination of qualities cannot easily be said to lack the others in the list (can a conscious, intelligent being really be said to lack personhood?)—and (c) the functional approaches in question are testing for this quality to see whether it can be ascribed to entities other than group agents, then we can use the same approaches as support for the thesis that personhood in List and Pettit's performative-relational sense also applies to other kinds of agents, and to artificial agents in particular.

I should mention, before we begin, that a similar approach has been suggested by Galliott (2015), who brings it to bear in dealing with the problem of responsibility in automated warfare. He turns in particular to the problem of "the supposed 'responsibility gap'—namely, the inability to identify an appropriate locus of responsibility" (ibid., 211)—and observes that it would not be too practical to address this problem from the perspective of the "classical accounts" (ibid., 224) of responsibility, with their emphasis on "free will and intentionality" (ibid.). Instead, we should take a pragmatic or functional approach to responsibility, which does not proceed from a concept of agency in thinking about responsibility (as List and Pettit do) but rather frames the discussion in terms of roles and norms. As Galliott notes in that regard, "both Daniel Dennett and Peter Strawson have long held that we should conceive of moral responsibility as less of an individual duty and more of a role that is actively defined by pragmatic group norms" (ibid.). This means that responsibility is ascribed not so much by identifying a cast of characters and asking what they did ("who did what?"), as by considering the norms by which they operate, the roles they play in following those norms, and the underlying rationales (i.e., what the aims and reasons are behind those norms and roles), very much in keeping with the performative-relational view just outlined, where an agent's capacities are conceived as capacities exercised in a social world governed by norms enabling different agents to interact without incident. Like List and Pettit's agency-centred approach, this functional approach "has the benefit of allowing *non-human* entities, such as complex socio-technical systems and the corporations that manufacture them, to be answerable for the harms which they often cause or contribute to" (ibid.; italics added), and like the functional approaches it fills variable *x*—the locus of responsibility in Galliott's case, the locus of personhood in our case—not by looking for characters who may fit the description but by focusing on *outcomes*:[13] Just as Galliott arrives at responsibility by *first* asking how a given set of norms and roles can yield a frictionless social environment and how we should reframe those norms and roles when accidents happen revealing that something is not working, and only *then* (if need be) looking at who or what was entrusted with those roles, so we arrive at personhood by asking what it is that an agent can do in a social environment, or whether it can perform in a socially congruent and beneficial manner.

Two differences between group and artificial agents

So far in this discussion we have considered how a conception of agency, responsibility, and personhood can be framed in such a way as to apply to group and artificial agents alike. This was done by abstracting from notions of agency, responsibility, and personhood that might be described as anthropocentric in virtue of their using the human being as the basic reference point for thinking about these questions. And by framing the discussion in this more abstract way—that is, by looking at what agents do or *can do*, rather than at whether they have a mind like ours or some other inherently human feature—we have been able

[12] Four such approaches are Hubbard (2011) (in which the *x* variable is personhood itself), Rothblatt (2014) (consciousness), Dennett (2013) (intelligence), and rights (Nussbaum 2006, 2011), and what they all have in common is that, in testing for a quality or property *x*, they do not ask us to imagine what it would be like to enter into the "mind" of the entity we think it might be ascribable to, but only ask us to consider whether this entity is functionally or operationally capable of acting consistently with what it means to have that quality or property.

[13] The approach "allows for the fact that agency develops over time and shifts the focus to the future appropriate behaviour of complex systems, with moral responsibility being more a matter of rational and socially efficient policy that is largely outcomes-focused" (Galliott 2015, 224).

to bring out some features that group and artificial agents importantly have in common. But no less important are the *differences* between group and artificial agents, and we will focus on two in particular: a structural difference (in "Structural difference" section), which is that a group agent is made up of individual agents, while an artificial agent generally is not; and a difference pertaining to the question of autonomy ("Two kinds of autonomy" section), which in a group agent is described by List and Pettit (2011, 76) as a *supervenient* autonomy, whereas in an artificial agent it is a *conferred but self-reinforcing* autonomy.

Structural difference

As previously mentioned, one of the difficulties we face in drawing analogies between group and artificial agents is that they are *structured* differently, which is to say that only group agents are made up of constituent (group or individual) agents, namely, the people who view themselves as part of the group. This suggests that no point of comparison can properly be established between structurally different entities. If we take a closer look at the concept of a group agent, however, we will see that it would be a mistake to draw that conclusion.

List and Pettit note that it's wrong to think of a group agent as simply a collection of its members, for if it were we could not ascribe agency to it. In their own words, a group agent is "a single entity and not the collection of its members," for this entity "is subject in its own right to the constraints of agency" (List and Pettit 2011, 194). Another way to say this is that a group agent results not from the *joint* action of its members but from their *corporate* action: In joint action, individuals work together toward a common goal, whereas in corporate action we start out with a group entity, and only when that entity is formed can it act the way its creators intended.[14] There is a deeper sense, then, in which a group agent can be said to form a single entity: Although it could not exist without its constituent members (which are therefore essential), it is not in these members that its agency lies, for this is not an agency that can be arrived at by summing up the agencies of all the group's members. And this shows that, while artificial agents may be structured in a different way than group agents, that difference does not stand in the way of our identifying relevant analogies between agents corresponding to those two descriptions.[15]

But, as mentioned, there is also a second argument that could be mounted in rejecting the idea of a parallel between artificial and group agents, in that they have different kinds of autonomy. In the next section, however, I will argue that even this argument does not stand up to scrutiny.

Two kinds of autonomy

The second argument against the view that group agents can suitably illustrate how artificial agents could be recognized as agents proper proceeds from the case of a corporation as a group agent in the law, and from the premise that a corporation is a fictitious (legal) person whose will is actually the will of its members (of those who own or run the corporation). Hence the conclusion that a corporation lacks autonomy. List and Pettit (2011) reject this view of group agency as fictitious and hence devoid of autonomy, and in the rest of this section I explain why and contrast their account of a group agent's autonomy with my own account of the autonomy of artificial agents, the latter account proceeding from the premise that artificial agents do have autonomy, for otherwise we wouldn't have to worry about their status as entities independent of their human developers and users or about their status as actors in the social world.

There are two different notions at work when we speak of autonomy in a group agent and in an artificial agent. Both are subject to limitations, as one might expect, so it is the *way* in which those limitations operate that we have to consider in fleshing out the two different types of autonomy in question.

The autonomy of a *group* agent is limited by its structure as an entity that could not exist without its members, especially in its operation, in that everything a group agent does must necessarily be done through its members: A group agent cannot do the things it does unless its members act in such a way as to bring about that result, and the group agent's action therefore cannot arise independently of that of its members (List and Pettit 2011, 64). And yet it does make sense to speak of a group agent as somehow acting independently, or as having a "mind of its own," and to explain that insight List and Pettit advance a

[14] Interestingly for our purposes, this very same reasoning was anticipated by Chief Justice John Marshall in the landmark case *Trustees of Dartmouth College v. Woodward* (1819), where it was applied to the concept of a business corporation: "From the nature of things, the artificial person called a corporation, must be created, *before* it can be capable of taking any thing. When, therefore, a charter is granted, and it brings the corporation into existence without any act of the natural persons who compose it, and gives such corporation any privileges, franchises, or property, the law deems the corporation to be *first* brought into existence, and *then* clothes it with the granted liberties and property" (italics added).

[15] Another parallel that can be drawn is between a group agent and a multi-agent system (MAS), a system composed of interacting individual agents (computer systems) acting to achieve a common goal (for an introduction to MASs, see Woolridge 2009). This parallel will not be addressed here because the artificial agents making up an MAS are different from the kinds of agents discussed in this paper.

supervenience thesis: A group agent's actions and attitudes (what the agent "does" and "thinks") *supervene* on those of its members, that is, they emerge on the basis of those latter actions and attitudes, and it is the group's structural design that determines how that happens.[16] To see this, we can take the authors' example of a democratically organized group agent versus a tyrannically organized one: What the group decides in the first case depends on what a majority of its members decide; what it decides in the second depends on what the dictator commands, regardless of what everyone else in the group thinks. This means that two groups may have the same attitude but may arrive at that attitude in different ways as a function of the group's procedural organization or structure.

So a group agent's range of action can be said to depend on two factors: on the spectrum afforded by its individual members—the spectrum of their attitudes—coupled with the way these attitudes are procedurally worked into a final outcome, that is, a final or emergent decision or attitude. But how does this emergence or supervenience amount to something like the group's autonomy?

The way List and Pettit tackle this problem is by putting forward the thesis of a "non-redundant realism" (List and Pettit 2011, 76). By this term they mean that group agents exist—they are real—and this reality cannot be reduced to that of their members: It cannot be "analyzed away" and thus collapsed into that of their members, and it is in this sense that the autonomy of group agents can be described as non-redundant.[17] However, it is not an ontological autonomy that we are looking at but an epistemological one: A group agent's existence, and hence its autonomy, is not something that's "out there" in the world—group agents do not exist as "hyper-realities" (ibid., 75)—so the best we can do is work out theories that will show us how their autonomy works.[18] But we have to be careful here. We saw earlier that no group agency could exist without the individual agencies of the people who make up the group: These people interact in complex ways so as to enable the group to act as a group with its own identity, and this suggests that we could unbundle these interlocking strands by tracing them to the individual actions and attitudes of the group's members. But this is precisely the error List and Pettit want us to avoid, arguing that if we embrace this methodological stance—a methodological individualism that would have us observe the group agent exclusively through the lens of its individual members—we will be prevented from seeing "the wood for the trees" (ibid., 76).[19]

We can see, then, that this epistemological group autonomy welds together two claims: On the one hand is the positive claim that by studying agents we can genuinely advance our knowledge about our social existence as framed by a complex of nontrivial interactions with our environment; on the other hand is the negative claim that this knowledge is not something we can gain just by studying the way the individual constituents of that social world behave (while neglecting all other factors). In fact, when it comes to figuring out the attitudes of group agents as supervenient on those of its members, we may be tempted to draw a straight line from one end to the other (from individuals to the group they form), but the relation may well be more complex than that, and as List and Pettit point out, there are three sorts of difficulties to account for that.

First, there is the difficulty of identifying the attitudes of individual members. Although we know that the group agent's attitudes are dependent on those of its individual members, it may be a challenge to identify and "count" those attitudes. For instance, we may know that the group agent has adopted an attitude based on what the majority of its members think, but we may not know how to determine the makeup of that majority.

A second difficulty arises when the group's attitudes depend not on those of its individual members but on the attitudes that different *sets* of members take to a complex of interconnected propositions. In order to overcome a difficulty of this sort, we would have to know which propositions count for a given group attitude, which sets of individual members count more than others, and so on—which, again, can become a challenge.

The third difficulty comes in when the group agent's organizational structure is unclear, as when decisions are taken by a (nonbinding) straw-vote procedure. In such cases, the group agent's attitudes still supervene on those of its individual members, but it becomes very complicated to date back the dynamics through which the group agent took the stance corresponding to those attitudes. In fact, supervenience in this case becomes bidirectional, since it may be

[16] As Tuomela (2011) has pointed out, the authors do not address the grounds of supervenience—causal, conceptual, or epistemic—and my own discussion of supervenience suffers form the same defect.

[17] There are a number of other theories that take this approach: See the table in List and Pettit (2011, 7).

[18] This is List and Pettit's way of striking a middle ground between two views of group agency which they term "emergentist" and "eliminativist": "Where emergentism makes group agents into hyper-realities, eliminativism makes them into non-realities" (List and Pettit 2011, 75). It is not entirely clear, however, how this middle-of-the-road view (epistemological autonomy) can be distinguished from the emergentist view, since List and Pettit use the same exact language to describe both: "From the emergentist tradition," they note, "it went without saying that group agents were agents in their own right, over and above their members" (ibid. 73); compare that with their own approach, on which "we must think of group agents as relatively autonomous entities—agents in their own right" (ibid., 77), thus defending "the idea that group agents can be agents over and above their individual members" (ibid., 78).

[19] Non-redundant realism is criticized by Sylvan (2012), arguing that group agents can be seen through the lens of a *redundant* realism.

the case that the individual members' attitudes supervene on the group agent's attitudes, which in turn supervene on the individual members' previous attitudes. This mutual feedback between individuals and the group means that supervenience may have an "evolving character" (ibid., 77), which compounds the difficulty involved in tracking the phenomenon.

I might comment, in this connection, that much of this mutual feedback between the group and its members consists of informal downward pressures which run from the former to the latter, and which tend to become increasingly pervasive and forceful as the group grows in size and complexity, as when we get to the level of society: These are the political and cultural forces that take shape through the power struggles which inevitably arise within any group of any considerable size, and they are such that the group may wind up forcing its own identity on its members, who may not necessarily view that identity as something they would otherwise espouse. These pressures raise three kinds of concerns. First, they are often inescapable, especially when the group within which they emerge is one its members cannot exit at will (examples here are the nation, the community, and even the corporation, if escaping the corporate culture and its pressures means looking for new employment in an unreceptive job market). Second, these pressures are persistent (as well as pervasive) and resistant to change (one need only think here of the laissez faire ideology which propelled the economic and industrial revolution in post-bellum America, and which still acts strong even to this day, despite the evident failures of the free-market system it is intended to support). And third, they make it difficult to analyze the dynamics that shape the relation between the group and the individuals within it, for they cannot easily be quantified or factored into any formal model.

So what happens in these situations, when facing these three kinds of difficulties in linking the group agent's attitudes to those of its members, is that the group agent acquires a degree of autonomy, an autonomy at once relative and epistemological, for on the one hand the group's attitudes are based on those of its members and are constrained by the latter (a bounded autonomy), and on the other we cannot fully derive the group's attitudes from the attitudes of its members (epistemological autonomy).

Having looked at group autonomy, we can consider the same characteristic (autonomy) in artificial agents. The autonomy of artificial agents is precisely what enables them to stand apart from previous "passive" technologies: This is true of "sense-think-act" technologies, in which the range of an agent's action is limited to that of the input they receive. That is different from the kind of behaviour we would recognize as properly autonomous, where an artificial agent can select not only the means through which to achieve its ends, but also the ends themselves. An overview of the literature in computer science suggests that an agent can be said to be autonomous if it can (*i*) learn from experience and act (*ii*) over the long course (*iii*) without the direct control of humans or of other agents (Laukyte 2012). This is still a *bounded* autonomy, to be sure—for it is *designed* into the agent, and so is an endowment the agent gets from its (human) "makers"—but it also gives the agent an increasing ability to bring its experience to bear on that autonomy so as to expand the kinds of ways in which it can successfully interact with its environment and bring about the desired end, and for this reason the agent's autonomy can be characterized as self-reinforcing.

We can now consider whether the difference between these two kinds of autonomy should prevent us from analogizing the two types of agency they describe. It is often thought that the magnitude of that difference does pose an obstacle in that regard, and the argument would typically run as follows: Whereas the autonomy of *artificial* agents, being bounded by design limitations, is too weak to enable such agents to qualify as responsible members of a socially networked environment—an environment framed by interactions governed by mutual expectations—the autonomy of *group* agents, being instead bounded by their organizational structure, places a much weaker constraint on an agent's autonomy (especially if this is a group agent whose members already enjoy full autonomy), and for that reason group agents are *not* prevented from qualifying as fully competent members of society. This I would call the minimum threshold argument—for it assumes there to be a minimum qualifying degree of autonomy an agent must possess in order to become eligible to participate in a social world—and I have two problems with that line of reasoning. For one thing, even granting that such a threshold can in fact be identified, what is presently limiting the autonomy of artificial agents is the technology we use to design and build them, and this is a practical impediment, not a necessary or principled one, so there is no reason to believe that those limitations cannot one day be overcome, especially considering that the autonomy of artificial agents is self-reinforcing.[20] And, for another thing, it seems uninformative to set a minimum degree of autonomy without considering the way in which that capacity is exercised: A group agent's autonomy is exercised through the group's members and through the procedures they use in coalescing their many voices into a single voice; an artificial agent's autonomy is exercised by the agent itself on the basis of the

[20] Consider in this regard the opinion expressed by the computer scientist and inventor Ray Kurzweil (quoted in Greenemeier 2010): "Machines will follow a path that mirrors the evolution of humans. Ultimately, however, self-aware, self-improving machines will evolve beyond humans' ability to control or even understand them."

design through which it operates. These different ways of exercising autonomy point to different capacities, and it is these capacities we have to take into account in judging whether an agent's autonomy makes that entity a properly *social* agent, and so an agent to which responsibility and personhood can be attributed.

So, having addressed two critical points of the parallel between group and artificial agents—arguing that neither their different organizational structure nor the different kinds of autonomy they embody are reasons for rejecting this parallel—I take up the implications of ascribing personhood and responsibility to artificial agents.

Implications of ascribing responsibility and personhood to artificial agents

It would be implausible to attribute personhood and responsibility to any kind of agent without working out what such an attribution would entail, especially considering that by conferring these two attributes we fashion a kind of agency at once social and moral: social in the sense that agents so characterized must relate to and interact with other agents; moral in the sense that any agent operating in such a relational world is bound to face choices about what to do vis-à-vis others, and these choices almost by definition invite moral considerations, and may even require a moral judgment about the best course of action in the situation at hand. As one might appreciate, a discussion so framed can easily expand out of proportion (covering anything touched by the word *social* or *moral*), and so in order to make it manageable I am going to restrict it to the question of the rights that can be claimed for artificial agents once it is recognized that they are endowed with personhood and can be held responsible.

As discussed in "Third condition of agency: responsibility" and "Fourth condition of agency: personhood" sections, agents of any kind (individual, group, or artificial) are ascribed personhood and responsibility on the basis of their capacities, or what they can do. This means that we have to design rights enabling them to exercise those capacities. I would accordingly call these "enabling rights," playing a role to similar to what John Rawls in his theory of justice as fairness called primary goods, defining them as "things that every rational man is presumed to want," in which regard he asks us to "assume that the chief primary goods are […] rights and liberties, powers and opportunities, income and wealth" (Rawls 1971, 62).[21] It is clear from the definition just offered that Rawls's primary goods cast a wide net, because in his theory the basis on which they can be ascribed is that of rationality ("every rational man"), whereas here the basis of ascription is that of an agent's capacities. So, on the one hand, enabling rights are similar to primary goods, in that both assume the existence of capacities or powers of reason whose exercise they are intended to enable, but on the other, enabling rights can be much more restrictive than primary goods, since the former are each tailored to specific capacities, whereas the latter "are things which it is supposed that a rational man wants whatever else he wants" (ibid., 92), so their design is essentially "one size fits all," considering that all men (or all agents, where we are concerned) are assumed to be rational.

Enabling rights, then, contain something of a paradox, because they can be both more *specific* than other types of goods or entitlements (in that they are each intended to support specific capacities) and more *universal*, in that they do not just apply to human beings but to any agent (human or otherwise) that can be thought of as rational. And that fact prompts two related questions connected with the two features just mentioned: How specific should enabling rights be? And how can we get them to all cohere? Or, more specifically, in virtue of their universal applicability, how can the enabling rights ascribed to nonhuman (group or artificial) agents be made coherent with those we recognize for human agents?

The first question cannot be addressed here in any exhaustive manner, because the specificity of each enabling right will depend on how specific the capacities are that we want to support on a case-by-case basis, and that judgment will depend on a variety of factors. Corporations, for example, have been recognized since the early nineteenth century as having the ability to sue and be sued, and that judgment—in *Trustees of Dartmouth College v. Woodward* (1819)—was based on a recognition of the increasingly important role of corporations in society, as well as on an interest in promoting economic growth and risk-taking in business.[22] So we see that a lot can go into the kind of reasoning required to answer that first question, which, as the example suggests, may very well involve an assessment of where we are in history and what kind of society we want to shape. Nor am I suggesting that just because corporations gained legal recognition in the nineteenth century as artificial persons having rights and duties, we should thereby take that status as justified, simply in virtue of its existence. To make that assumption would be tantamount to extracting normative conclusions from factual premises

[21] Rawls would later be criticized by Habermas (1995, 114) for assimilating rights and liberties to goods—which are more like property, or things you *own*—but that is a matter that would take us on a long detour, so it cannot be taken up here.

[22] On the historical context in which that judgment and recognition came to be, see Friedman 2005, 136–37. For a broader discussion of corporations as rights-holders, see Clements (2012).

(thus coming up against Hume's is-ought problem). Rather, as can be appreciated from the foregoing remarks, any ascription of rights to any sort of agent requires a broad assessment of the *reasons* why those rights ought to be ascribed: These reasons are inevitably going to be normative, and they inevitably have to extend beyond a recognition that the agent in question is endowed with certain capacities; at a minimum, on the performative-relational approach I am putting forward, we must consider how an agent so endowed (with a set of capacities and corresponding rights) is going to interact with other agents in the broader social and political environment.

One general remark *can* be made, however, in addressing the first question. It is that the broader we make the capacities worthy of protection, and so the broader we make the enabling rights supportive of those capacities, the more we set the *subjects* of those rights (the agents recognized as rights-holders) on an equal footing, in that agents with equal capacities are assumed to have equal rights. So, for example, there is a risk in choosing a broad concept of rationality as a basis of ascription, because the broader the concept, the more inclusive it will be, and the more it will apply to human and nonhuman agents alike. This is the second feature of enabling rights (their universal applicability), and it takes us to the second question, that of the coherence or coexistence of rights.

This second question can be answered in a more systematic way than the first. To see this, we first have to set up a contrast with the view espoused by List and Pettit (2011, 180), for whom there is a clear criterion for ascribing rights to group agents: These rights should be recognized only insofar as that works out to the benefit of human beings. This is what I am calling the anthropocentric viewpoint, and it is a perfectly reasonable approach where group agents are concerned. After all, it is we—as natural human persons—who create or give rise to group agents, and the latter wouldn't exist without us, whereas the converse relation *would* seem to hold: Individual agents, it would seem, can exist without group agents, or at least they can be thought of as existing without requiring group agency as a background condition. In fact, no person can live in complete isolation: Everyone has to be part of some group or community at some point in their lives or in some capacity. If we accept that, then we can begin to see that what appears to be a one-way relation between individual and group agents is actually a two-way relation, and we saw a bit of that when we considered the third difficulty that List and Pettit (ibid., 77) point out in as a stumbling block in any attempt to reduce a group's attitudes to those of its members.

If we develop that observation, we arrive at a viewpoint I would call inter-relational, in distinction to the anthropocentric viewpoint. From an inter-relational viewpoint, we see that group agents depend on individual agents as much as the latter depend on the former; or—otherwise stated, going back to List and Pettit's third difficulty ("Two kinds of autonomy" section)—group attitudes are shaped by individual attitudes as much as the latter are shaped by the former. This appreciation should encourage us to embrace a different perspective, from which instead of asking, How do humans stand to benefit from an ascription of rights to nonhuman agents? we ask, What kind of social environment are we shaping by making such an ascription, and is it a kind of environment we would like to live in?The two questions may very well lead to the same answer in any given case, to be sure, but they frame the problem differently, for on the one hand we set ourselves up for thinking about rights in terms of what benefits us as humans or as individuals, whereas on the other we can step back and take a broader view not *closely* focused on human welfare or on what benefits us in the short term.

Now, my contention is that if we design enabling rights on the basis of the competence approach and then view those rights from what I am calling the inter-relational viewpoint, we have a systematic way by which to address the question of how to construct a framework in which human and nonhuman rights can form a coherent whole. We do so by viewing ourselves not as the reason why our social envi-ronment exists but as an essential part of that environment, the idea being that what enables that environment to thrive enables *us* to thrive as well. This broad conception can be applied specifically to artificial agents in two stages: First, we recognize these agents as having agential capacities (or sets of capacities on which basis they can be counted as responsible agents qua persons), and second, we recognize that these agents interact with us within a society that sus-tains us all. And as List and Pettit (2011, 5) themselves point out, the reason why we should consider an artificial agent an agent proper is that if we do so "we can interact with it, criticize it, and make demands on it, in a manner not possible with a non-agential system." So once we start thinking that way (at this second stage), we have already embraced the inter-relational stance, whose point is to show how we can work toward "a global society in which all persons, on the basis of their capacity of thought and feeling, can participate as equal citizens, control their own affairs and achieve their fullest potential, regardless of the characteristics of their bodies" (Hughes 2004, 82).

So, in summary, the argument I am making is that once we can recognize an artificially intelligent system as having capacities that make it a rational and responsible agent endowed with personhood, and once we recognize that an agent with those capacities is at the same time an interactive entity that is going to inhabit and shape our social environment, then we have the premises on which to claim that that agent can be recognized as having enabling rights corresponding to those capacities, not only because we can

engage in practical reasoning with an agent so described ("we can interact with it, criticize it, and make demands on it," as just noted) but also because the environment shaped by such agents is a networked environment whose members are interdependent and owe their existence to it. The argument is not that we *have* to build such artificially intelligent systems—that trend is already afoot: It is the direction we are already heading in—but that once we do have those systems and they have the requisite agential and inter-relational capacities, we have a basis on which to make them an integral part of our social environment, recognizing them as having the enabling rights and corresponding duties that go with those rights.

Closing remarks

An important strand of the argument I have developed for recognizing artificial agents as members of the social world we share with them has rested on an analogy between artificial and group agents. The debate on group agency, responsibility, and personhood is ongoing, and much of it is focused on the legal personhood of corporations [see, for instance, Westra (2013) and Hartmann (2010)]. That is why I have picked up the question of corporations as subjects of rights, while also referring to the historical process through which corporations have been recognized as legal persons. The point of that discussion and analogy was not that since corporations have been recognized as having that status, so should artificial agents. The point was rather to explore what the reasons are on which basis group agents, such as corporations, can, qua agents, be recognized as responsible members of our social and political environment, and whether the same reasons might apply to artificial agents. The account of agency I have laid out is intended to offer a framework within which to answer that question, and so the question of whether and on what basis artificial agents can play a role as members of the increasingly networked environment we are building.

So while I do argue for recognizing artificial agents as members of a social world, I couch that argument within an account of agency meant to clarify what is at stake and how the whole question might be approached. This is just one approach in a debate that has engaged philosophers, scientists, theologians, lawyers, and social scientists in an effort to work out a range of related issues in the budding field of roboethics, concerned with the moral considerations that we should be making in designing and using robots.[23] Owing to the wide use of robots and artificial intelligence, the issues span from those of personal identity (the enhanced self) and the interpersonal sphere (companion robots) to the socioeconomic (robot displacement of human workers) and national security [the use of robots in the military: Galliott (2015), Sparrow (2007), and Krishnan (2009)]. My own discussion looks out a bit further into the future by anticipating a world in which the technology will have been built that makes fully intelligent artificial agents already a reality, and in this scenario I ask how our relation to these agents should be framed.

I would like to close the discussion by pointing out two ways in which my line of thinking can be developed going forward. One way is to take a historical view in comparing group and artificial agents and arriving at a fuller understanding of both. A hint of how the historical view can come into the picture was offered earlier on (in "Implica-tions of ascribing responsibility and personhood to artificial agents" section) in connection with the question of the corporation, where it was briefly discussed how its role in society developed over time and how a variety of consid-erations may go into a judgment about that role and what it should be. I am thus suggesting that it may prove illumi-nating to consider not only the theory but also the devel-opment of agents over the long term. Where corporations are concerned, their history reveals that the early ones established under UK and US law were quite different from the currently operating ones. The early corporations, for example, could only operate on national territory and could not control other corporations. These conditions were loosened over time, and now, in *Burwell v. Hobby Lobby* (2014), closely held for-profit corporations have been found to have the right to assert religious convictions just as individuals can. From a historical perspective we can thus see a broad trend toward looser and looser restrictions on what a corporation can do under the law, or what capacities a corporation has and what rights ought to go along with those capacities. So, if corporations have seen this development, and the previously developed analogy between group (corporate) and artificial agents holds up, then we can begin to consider the ways in which even artificial agents can be envisaged to follow a similar path: This is something we can do by exploring the reasons that may be adduced in making such an argument and figuring out on that basis what that could mean for the legal regulation of artificial intelligence.[24]

[23] For an overview of the roboethics debate see, for instance, Lin et al. (2012).

[24] The important point here is the emphasis on *reasons*: As previously mentioned, I am not suggesting that because history or the law evolved as it did in regard to corporations, then we should mimic the same line of development in dealing with artificially intelligent agents. Rather, I am saying that the analogies that group agents (and corporations among them) can be shown to have to artificial agents warrant an investigation aimed at exploring whether the justifications for one development (in the past) are sound and might also justify another development (in the future).

The second avenue I'd like to suggest takes its cue from its interpretation through the inter-relational stance. It does so by working on the idea of the social and natural envi-ronment that comes into focus once we take that stance. It was previously noted that the question of ascribing rights to nonhuman agents on the basis of their capacities can be framed in either of two ways: We can ask how we humans stand to benefit *directly* from such an ascription of rights, or we can ask how we can improve our lot through the *environment* we forge through the same ascription. And it is this latter framing of the question that I believe we should stress. This framing connects us to the idea of environmental ethics, envisioning a framework of rights inclusive of all entities, a framework that cannot be com-plete without addressing the rights of the environment as the foundation of our own wellbeing. This very idea was pushed even further in the 1970s by Naess (2010) in a conception he calls Deep Ecology, on which nature is inherently valuable regardless of whether it is useful for humans or animals. And this, too, is a conception we can turn to in thinking about artificial entities, for it shows us a way to frame the question of their agency and of the rights ascribable to them without having to invoke sen-tience-based categories.

Finally, I should note that the inter-relational framing of the question of the rights ascribable to nonhuman agents is consistent as well with the cosmopolitan vision advocated by Martha Nussbaum in urging that we become "citizens of the world" (Nussbaum 1997, 52). In making that argu-ment, "Nussbaum takes us back to the Stoics and their image of concentric circles of affiliation, going from self and family out to the nation and finally to the widest circle, which embraces all of humanity" (Fischer 2007, 153).[25] This is an image that comes to us by way of the Stoic philosopher Hierocles, who thought that it was our task "to 'draw the circles somehow toward the center,' moving members of outer circles to the inner ones" (Nussbaum 1994, 342) in a process where "the ultimate aim would be to treat all men as our brothers" (Sandbach 1989, 34). I would therefore suggest, in keeping with the Deep Ecology previously mentioned, that in this process we can draw an even larger circle extending beyond humanity so as to take in not only the whole of humanity but also nonhuman entities like artificial agents, and we can do so without necessarily subscribing to a Stoic ethic, for we have independent support for that move on the competence approach as outlined in this article.

I believe that if we can take all this into account, we will have a basis on which to address further questions about the way we ought to deal with artificial agents. One set of questions revolves around the people and the processes to which we should entrust the decisions we make about artificial agents, that is, (a) *Who* should make these decisions—intellectuals (ethicists, philosophers, and the like), technicians (engineers, developers, and the like), legal professionals (judges, lawyers, jurists), policymakers, or all of the above?—and (b) *How* should the decision-making process be organized and how many voices should be brought into the conversation? And another set of questions, perhaps further afield, revolves around artificial agents themselves as full participants in our social environment, where we can start to think about the "life" of artificial agents, as well as about their health, emotions, thoughts, and the like, asking, for example, What is it for an artificial agent to have a full life according to its capacities and its role in our social environment? These questions may be somewhat esoteric at this point, but they should not be discounted, and the approach I have outlined can offer a framework within which to address them.

[25] For a critique of Nussbaum's cosmopolitanism, see Ayaz Naseem and Hyslop-Margison (2006).

Acknowledgments I would like to thank Filippo Valente for copy-editing this article and making a few helpful suggestions along the way. I would also like to thank the anonymous reviewers pointing out several ways in which the argument could be improved.

References

Ayaz Naseem, M., & Hyslop-Margison, E. J. (2006). Nussbaum's concept of cosmopolitanism: Practical possibility or academic delusion? *Paideusis, 15*(2), 51–60.

Briggs, R. (2012). The normative standing of group agents. *Episteme, 9*(3), 283–291.

Burwell, Secretary of Health and Human Services, et al. *v. Hobby Lobby Stores, Inc*., et al. 2013. 573 U. S. (2014). http://www.scotusblog.com/case-files/cases/sebelius-v-hobby-lobby-stores-inc/. Accessed 5 Oct 2014.

Clements, J. D. (2012). *Corporations are not people: Why they have rights that you do and what you can do about it*. San Francisco, CA: Berrett-Koehler Publishers.

Dennett, D. C. (2009). Intentional systems theory. In B. McLaughlin, A. Beckermann, & S. Walter (Eds.), *The Oxford handbook of philosophy of mind* (pp. 339–350). Oxford: Oxford University Press.

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. New York: W. W. Norton & Company.

Dietrich, E. (2011). Homo sapiens 2.0: Building the better robots of our future. In M. Anderson & S. Anderson (Eds.), *Machine ethics* (pp. 531–541). Cambridge: Cambridge University Press. Emerson, R., & Hardwicke, J. W. (1997). *Business Law*. Hauppauge, NY: Barron's educational series.

Fischer, M. (2007). A pragmatist cosmopolitan moment: Reconfiguring Nussbaum's cosmopolitan concentric circles. *The Journal of Speculative Philosophy (new series), 21*(3), 151–165.

Friedman, L. M. (2005). *A history of American law*. New York, NY: Touchstone.

Galliott, J. (2015). *Military robots: Mapping the moral landscape*.

Greene, J. D., et al. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.

Greenemeier, L. (Ed.). (2010). 12 Events that will change the world. *Scientific American, 302*(6), 36–50.

Habermas, J. (1995). Reconciliation through the public use of reason: Remarks on John Rawls's political liberalism. *The Journal of Philosophy, 92*(3), 109–131.

Hartmann, T. (2010). *Unequal protection: How corporations became "people"—And how you can fight back*. San Francisco, CA: Berrett-Koehler Publishers Inc.

Himma, K. (2009). Artificial agency, consciousness and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19–29.

Hohfeld, W. N. (1917). Fundamental legal conceptions as applied in judicial reasoning. *Faculty Scholarship Series.* Paper 4378. http://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=5383&context=fss_papers. Accessed 10 Sept 2016.

Hubbard, P. E. (2011). "Do androids dream?": Personhood and intelligent artifacts. *Temple Law Review, 83*, 404–474. Hughes, J. (2004). *Citizen cyborg*. Cambridge, MA: Westview. Illes, J. (2005). *Neuroethics: Defining the issues in theory, practice and policy*. New York: Oxford University Press.

Jones, P. (1994). *Rights*. London: Macmillan.

Krishnan, A. (2009). *Killer robots: Legality and ethicality of autonomous weapons*. Farnham: Ashgate.

Laukyte, M. (2012). Artificial and autonomous: A person? In G. Dodig-Crnkovic, A. Rotolo, et al. (Eds.), *Social computing, social cognition, social networks and multiagent systems social turn (SNAMAS 2012)* (pp. 73–78). Birmingham: AISB.

Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2012). *Roboethics: The ethical and social implications of robotics*. Cambridge, MA: The MIT Press.

List, C., & Pettit, P. (2008). Group agency and superve nience. In J. Hohwy, J. Kallestrup (Eds.), *Being reduced: New essays on reduction, explanation, and causation* (pp. 75–92). New York: Oxford University Press.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford: Oxford University Press. Naess, A. (2010). *The ecology of wisdom: Writings by Arne Naess*. Berkeley: Counterpoint.

Nussbaum, M. (1994). *The therapy of desire: Theory and practice in hellenistic ethics*. Princeton, NJ: Princeton University Press. Nussbaum, M. (1997). *Cultivating humanity: A classical defense of reform in liberal education*. Cambridge, MA: Harvard Univer-sity Press.

Nussbaum, M. C. (2006). *Frontiers of justice: Disability, nationality, species membership*. Cambridge, MA: Harvard University Press. Nussbaum, M. C. (2011). *Creating capabilities: The human development approach*. Cambridge and London: The Belknap Press of Harvard University Press.

Pettit, P. (2001). *A theory of freedom: From the psychology to the politics of agency*. Cambridge: Polity Press.

Pettit, P. (2007). Responsibility incorporated. *Ethics, 117*, 171–201. Rawls, J. (1971). *A theory of justice*. Cambridge, MA: The Belknap Press of Harvard University Press.

Rothblatt, M. (2014). *Virtually human: The promise—And the Peril—Of digital immortality*. New York: St. Martin's Press. Sandbach, F. H. (1989). *The stoics* (2nd ed.). Cambridge: Hackett. Singer, P. S. (2009). *Wired for war: The robotics revolution and conflict in the 21st century*. London: Penguin Books.

Singer, A. E. (2013). Corporate moral agency and artificial intelligence. *International Journal of Social and Organizational Dynamics in IT, 3*(1), 1–13.

Solum, L. B. (1992). Legal personhood for artificial intelligences. *North Carolina Law Review, 70*, 1231–1287.

Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy, 24*(1), 62–77.

Sylvan, K. L. (2012). How to become a redundant realist? *Episteme, 9*(3), 271–282.

Trustees of Dartmouth College v. Woodward. 17 U.S. 518 (1819). http://caselaw.lp.findlaw.com/scripts/getcase.pl?court=US&vol=17&invol=518. Accessed 29 Oct 2014.

Tuomela, R. (1984). *A theory of social action*. Dordrecht: Kluwer. Tuomela, R. (2011). Review of Christian list and Philip Pettit *Group agency: The possibility, design and status of corporate agents*. *Notre Dame Philosophical Reviews.* http://ndpr.nd.edu/news/ 27604-group-agency-the-possibility-design-and-status-of-corpo rate-agents/. Accessed 11 Nov 2015.

Westra, L. (2013). *The supranational corporation: Beyond the multinationals*. Leiden: Brill.

Woolridge, M. (2009). *An introduction to multiagent systems* (2nd ed.). Chichester: Willey.