ARTEFACTUAL AGENCY AND ARTEFACTUAL MORAL AGENCY

Deborah G. Johnson and Merel Noorman,

University of Virginia

dgj7p@virginia.edu; mn7m@virginia.edu

Abstract This chapter takes as its starting place that artefacts, in combination with humans, constitute human action and social practices, including moral actions and practices. Our concern is with what is regarded as a moral agent in these actions and practices. Ideas about artefactual ontology, artefactual agency, and artefactual moral agency are intertwined. Discourse on artefactual agency and artefactual moral agency seems to draw on three different conceptions of agency. The first has to do with the causal efficacy of artefacts in the production of events and states of affairs. The second can be thought of as acting for or on behalf of another entity; agents are those who perform tasks for others and/or represent others. The third conception of agency has to do with autonomy and is often used to ground discourse on morality and what it means to be human. The casual efficacy and acting for conceptions of agency are used to ground intelligible accounts of artefactual moral agency. Accounts of artefactual moral agency that draw on the autonomy conception of agency, however, are problematic when they use an analogy between human moral autonomy and some aspect of artefacts as the basis for attributing to artefacts the status associated with moral autonomy.

Introduction

This chapter takes as its starting place that artefacts, in combination with humans, constitute human action and social practices, including moral actions and practices.[1] Identifying and differentiating the entities that make up the world is the work of ontology, and the ontology implicit in ordinary language and informal thought

[1] This material is based upon work supported by the National Science Foundation under Grant No. 1058457. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

seems to presume three fundamental kinds of entities: natural, human, and artefactual. Artefacts are individuated as entities through mental acts that separate human-fashioned materiality from naturally occurring materiality and from human activity and meaning. This ontology is the backdrop against which questions of agency typically arise. That is, having divided the world into categories of things, scholars and theorists ask where agency is to be found. Generally, humans are presumed to have agency while the agency of nature and artefacts are in dispute (each in distinctive ways). And, once the question of agency is raised, the further question of moral agency comes into focus. If artefacts have agency, why would they not have moral agency?

Ideas about artefactual ontology, artefactual agency, and artefactual moral agency are intertwined. In order to get a handle on the debate about artefactual moral agency, artefactual ontology and artefactual agency must first be addressed. After making the case for artefacts to be understood as components in larger sociotechnical systems, we distinguish three conceptions of agency: causal efficacy, acting for, and moral autonomy. We then take up the issue of artefactual moral agency arguing that conceiving of artefacts as moral agents can be productive when it refers to the causal efficacy of artefacts or to the tasks that have been delegated to artefacts by humans. However, understanding artefactual moral agency in terms of moral autonomy is problematic.

Artefactual Ontology

Artefacts are defined and generally understood to be human-made material objects. Although, as already suggested, the ontology embedded in our language and ways of thinking and speaking presumes three kinds of entities, when pressed, most of us acknowledge that the things in these categories overlap. We make statements of the following type: 'humans are part of nature'; 'artefacts are made by humans'; 'nature constrains what humans can do'; and 'artefacts are made by manipulating nature.' So, although the three types of entities are distinguished, they are inseparable; they are incomprehensible separately. Artefacts do not exist without humans making them; humans are part of the natural world; nature is understood as the 'stuff' from which humans come. This inseparability means that artefacts are never just artefacts.

Particular artefacts are individuated as entities by mental acts that draw ontological lines. To comprehend the significance of line drawing, consider the refrigerator of one of the authors.[2] Deborah's refrigerator is an artefact, that is, the chunk of plastic and metal that sits in her kitchen is an artefact. Some might even

[2] To be sure, refrigerators are more complex than, say, forks and bowls or hammers, but a complete typology of artefacts would take too long to introduce here. Later the distinction between computational and non-computational artefacts will be addressed.

say that her refrigerator is an autonomous entity (artefact) because it maintains its internal temperature "on its own". The thermostat in her refrigerator detects the temperature and signals other components of the refrigerator to change states so as to raise or lower the internal temperature. In this respect Deborah's refrigerator might be described as an autonomous agent acting on her behalf. Admittedly, her refrigerator's so-called agency does make a difference in her life. It allows her to conveniently eat and drink all kinds of things that might otherwise spoil.

The problem with characterizing Deborah's refrigerator as an artefact (and especially as an autonomous artefact) is that it only keeps her food cool when it is plugged into an enormously complicated power grid. Indeed, her refrigerator can easily be understood not to be an entity in itself but to be (merely) a component in a larger *technological system*. It is connected to a complex of artefacts – the electrical socket in her kitchen, the wires that run through her house and out to the street, the power station maintained by a company named Dominion Virginia Power. Going the other way, that is, breaking the rectangular chunk of metal and plastic into its component parts also suggests a technological system, for the rectangular chunk of materiality sitting in Deborah's kitchen is itself a combination of many different artefacts – a motor, vents, wires, metal parts, plastic shelves, etc.

The fact that Deborah's refrigerator is a technological system (meaning that it is multiple chunks of metal and plastic) and that it is a component in a larger technological system, is, however, only part of the story. A refrigerator only works as a refrigerator when human beings behave in certain ways. Deborah has to plug the rectangular chunk of metal and plastic into a socket; an electrician had to lay wire connecting the socket to a power grid; all the people working for Dominion Virginia Power have to come to work each day and do their jobs. In fact, the institutional arrangements constituting the power station are an enormous feat of human social organization and cooperation. In addition to those who work at Dominion Virginia Power are many other human beings and especially Deborah. She is needed to buy food, to open and close the door to put the food in and take it out. Importantly, she has to pay her utility bill (or else the Dominion Power will disconnect her refrigerator from the grid). Other humans are involved as well, for she could not buy food that needs refrigeration unless grocery stores carry it, and this in turn requires that trucks and airplanes bring refrigerated foods from far off lands to her grocery store. So, Deborah's refrigerator is not just a technological system; it is a *sociotechnical* system.

Where does the entity that Deborah calls her refrigerator begin and end? It seems that we have collectively and conventionally drawn a line. We have decided we will count the rectangular chunk of plastic and metal (the artefact) that sits in her kitchen as 'a refrigerator.' We have decided to leave on the other side of the line (outside of the concept of refrigerator) such components as the electrical grid to which her refrigerator must be connected, all the people who maintain the electrical grid, Deborah who must open and close the door to put in and take out food, the grocery stores, the trucks that deliver items to the grocery store, the global

trade markets that bring foods needing refrigeration to her grocery store, and so on.

The ontological line that we draw delineates Deborah's refrigerator as an artefact. In doing this we mentally and selectively extract it from the world in which it functions and has meaning; we disconnect it from all the other entities (human and material). In doing so, we make it 'something'; we think of it as something in itself. The mental act of thinking of it as an artefact blinds us to all of the activity behind the scenes (offstage), activity that makes her refrigerator function in the way she has come to expect. That her refrigerator is a sociotechnical system, that it achieves its results through a combination of human and non-human activity becomes something that we must work to see, against the backdrop of the artefactual ontology implied in ordinary language. It is not that we can never understand the connections among parts; obviously we can. The point is that the ontology draws attention to some of what is going on and directs attention away from other things that are going on.

Some might say that what we have just explained is the difference between artefacts and technology. Artefacts are material objects; technologies are sociotechnical systems. Artefacts are components in sociotechnical systems. This framework would seem to allow us, then, to ask what part artefacts play in sociotechnical systems and to ask whether the artefacts have agency.

Some of those who are particularly focused on computational artefacts might accept the distinction between artefacts and technology but insist that computational artefacts are different – because they are more autonomous or because they are autonomous in a distinctive way. This difference might then mean that computational artefacts can have agency when other artefacts do not. Consider an automatic pilot system. We can easily think of an automatic pilot software system as an agent acting on behalf of humans. It does many of the tasks that human pilots used to do and still do (when the automatic pilot is turned off). Of course, the reason the automatic pilot can control the airplane is because it was designed to do so and has been delicately connected to various other components of the airplane. So, whether or not the automatic pilot is autonomous is not a simple matter and whether its agency is different from other artefacts is not obvious.

Automatic pilots function only in combination with humans. In the design of automatic pilot systems, humans decide when and how the automatic pilot takes control of the airplane. Indeed, how automatic pilots work with humans can vary. Most automatic pilots are designed so that they take over control of the plane only when humans tell them to (e.g., when a human flips a switch). Of course, they could be designed so that they take control independent of any immediate human activity (that is, when they receive signals from other artefacts, internal to the airplane) or when humans at remote locations do something. And, of course, humans could decide to assign more and more of this decision making (i.e., when to go into automatic pilot control) to the technological components. The point, however, is that the automatic pilot, like the human pilot, *only* functions when it is a component in a larger sociotechnical system. To refer to the automatic pilot as an agent is

then to draw a line around a particular part of that system. One might do this in order to draw attention to its behaviour or its significance apart from or perhaps in interaction with the other components in the system.

Much of this is well-trodden territory. STS scholars have been especially concerned with how and why lines (boundaries) are drawn between humans and machines. For example, Suchman (2001) writes: "I take the boundaries between persons and machines to be discursively rather than naturally effected, and to be always available for refiguring." That lines are drawn between humans and machines (or artefacts) goes hand-in-hand with lines being drawn around artefacts.

The lines drawn are not innocent, they have real social and material consequences (Barad, 1996). Lines are drawn to make sense of the world, to facilitate practices, to give meaning, to achieve tasks. Delineating 'refrigerator' as an artefact containing shelves, doors, freezing elements, wires, nuts and bolts, rather than as a sociotechnical system, may make it easier to talk about a particular part that can be pointed to, moved, chosen, sold, etc. Yet, alternative ontologies are possible and can make a difference in what is seen and understood. For example, in Chapter 3 of this volume, Introna argues for a new ontology that better reflects the co-constitution of artefacts and humans. He emphasizes how each part is what it is because of other parts and traditional (human-artefact) line drawing works against our being able to notice this.

Artefactual Agency

Given the intricacies of delineating artefacts, what does it mean to say that artefacts have agency? It seems odd, on the face of it, that the question of agency would be raised with respect to 'things' that have been mentally constructed as chunks of materiality. Why draw lines around a chunk of materiality, extracting 'it' from a dynamic socio-material whole, and then ask whether (or proclaim that) the delineated chunk has agency?

One plausible answer to this question is that attributing agency to artefacts draws attention to (emphasizes, punctuates, makes visible) the role and significance of chunks of human-fashioned materiality in constituting the human world. This would, in turn, draw attention to the importance of decisions about fashioning and deploying those chunks of materiality. Another plausible answer (not unrelated to the first) is that thinking about artefacts as having agency is a useful way of understanding those chunks; it allows us to see aspects of materiality that we might not otherwise notice. For example, thinking of artefacts as having agency might allow us to see that they are far from inert, passive or neutral.

Although both answers seem plausible, more seems to be at stake in the discourse around artefactual agency. Attributions of agency to artefacts seem to do more than claim that agency is a useful concept. Indeed, since humans use language in complex, creative, and often fanciful ways, attributions of agency to arte-

facts may have a variety of functions or meanings or illocutionary uses. This is all the more likely because agency is such an unclear concept (Lee and Brown, 1994). Agency generally refers to the ability or capacity of an entity to act in the world. However, as explained below, many different conceptions of agency have been articulated and used in particular contexts.

Ironically, although unclear, agency is an important concept. It anchors many important discourses – moral discourse; discourse about what it means to be human; discourses about human relationships with animals, the earth, and transcendental beings; discourse about human rights and discourses about power and accountability. Indeed, the fact that agency is such an important concept and that it is so poorly understood may be connected; that is, its blurry meaning may facilitate use of the concept of agency in so many different contexts.

Three Conceptions of Agency

Discourse on artefactual agency seems to draw on at least three different conceptions of agency. The first has to do with causality. Many attributions of agency point to the *causal efficacy* of artefacts in the production of events and states of affairs. The second conception of agency might be thought of as *acting for* or on behalf of another entity: agents are those who perform tasks for others and/or represent others. The third conception of agency has to do with *autonomy*; agents are entities with the ability to think, decide, and intend, and to act accordingly. Distinguishing these three conceptions of agency is key to understanding discourse about artefactual agency and artefactual moral agency. Problems arise when one conception is conflated with another.

Causal Efficacy

If one wants to explain how the world got to be the way it is or how it currently works, or if one wants to shape the world of the future, thinking about causality seems unavoidable. One need not be a determinist to accept that things happen because of things that came before; one does not have to be a determinist to recognize that to get to a future state, events and changes will have to occur between now and then. Much of the discussion of artefactual agency – explicitly or implicitly – seems to have to do with the causal efficacy of artefacts in bringing about states of affairs. The design and availability of artefacts facilitate and constrain human behaviour. Whether the artefacts operate independently in time and space from humans (as in the case of thermostats controlling the temperatures in refrigerators) or they are deployed via direct human control (as when a person presses the trigger on the gun), artefacts make a difference in what humans do and what

happens in the world. The availability and design of particular artefacts affects how humans think, act, and organize themselves.

Causal efficacy is at least part of what is claimed by many STS theorists when they refer to the agency of artefacts. STS scholars have emphasized the affordances and constraints of artefacts and the contributions they make to outcomes, i.e., ongoing states of affairs, predicted futures. Actor Network Theory (ANT) is a good case in point (Law and Hassard, 1999); in treating nature, artefacts, and humans symmetrically, ANT acknowledges the causal contribution of all three in technological outcomes. Of course, ANT theorists insist – in effect – that the causal efficacy of each node in a network is dependent on other nodes. For this reason, it may be more accurate to say that ANT draws on the notion of causal efficacy, but is not reducible to it. Notice that in being ecumenical about artefacts, humans, and nature, ANT denies any special (a priori) status for any category of entity.

Although artefacts have causal efficacy, it is important to remember that artefacts only have causal efficacy in combination with humans. Humans design and deploy artefacts. Artefacts have meaning and function in relation to humans and human endeavours. We can draw lines around artefacts and we can extend or interpret the concept of agency so that we think and speak of artefacts as acting, but in doing so, we run the risk of pushing out of sight the human activity that is intertwined with the non-human activity. Suchman refers to this human activity as the 'offstage' or 'behind the scenes' activity that we may not notice but without which machines/artefacts do nothing (1998). Remember the refrigerator can only keep Deborah's food fresh if she plugs it in, pays her electricity bill, buys the food at a grocery store, and only if Dominion Virginia employees come to work every day, etc. All of this activity may be invisible when we think of the refrigerator as an agent.

Acting For

Human activity is explicit in the second conception of agency. Discourse on the agency of artefacts often seems to involve the idea that agents are those who perform tasks for humans or act on behalf of humans. In legal contexts, agents are those who are authorized to negotiate on behalf of a principal, as in the case of real estate agents or literary agents (Heath, 2009). Here agency involves representation, though the representation involves the agent using his or her expertise to perform tasks for the client. Latour's analysis of artefacts in "Where are the missing masses?" (1992) draws on this type of agency together with causal efficacy. He treats artefacts as if their role is to replace human actors; that is, artefacts do the (causally efficacious) work that human actors used to do or would have to do were the artefacts not there. He describes this as machines being delegated part of *the program of action*. The mechanical door groomer replaces the human that stood in

front of the door and opened and closed it as people came along; the traffic light replaces the police officer standing in traffic and using hand signals. These artefacts perform delegated tasks both on behalf of those who deployed (situated) them and those who encounter them. Regulators and engineers placed the traffic light at a crossroad to act on their behalf in enforcing moral behaviour from drivers, cyclists and pedestrians. They delegate this task or act by inscribing their intentions in the design of the traffic light; the traffic light expresses these intentions in signalling people when to stop and go. Similarly the mechanical door groomer was situated by architects, builders, and building owners to direct people to a particular place and to assist them in entering and leaving the building. It is thus not just the causal efficacy that is important here; it is the idea that artefacts affect human action through the delegated intentions inscribed in their design.

In a similar way, some computer scientists use an 'acting for' conception of agency to describe interactive software programs that accomplish tasks on behalf of their users, such as finding relevant news items online. Such *artificial agents* may perform decision-making tasks as well as negotiate with other agents. Computer scientists use the term agent here to mark a difference with other kinds of computer technologies. That is, artificial agent programs are different in that they are able to learn their users' interests, habits and preferences and use this information as they roam the Internet and carry out tasks for the users.

The 'acting for' conception of agency draws on a metaphor. Calling the mechanical part of the doorframe a door groomer makes an analogy with the human door groomer; calling an autonomous vacuum cleaner a housekeeper makes an analogy with the human housekeeper. Similarly, computer scientists and others refer to computer programs as software agents *as if* they acted on our behalf in the way that a servant or a hired worker might.[3]

Metaphors are more than ornamental devices or tools of persuasion in rhetoric. They are useful in making the unfamiliar, familiar. Metaphors help us to understand and make us comfortable with what might otherwise seem too complicated or alien. They allow us to see aspects of a thing that might otherwise be opaque. For example, referring to certain kinds of software as software agents that work on our behalf may help us to understand and explain what a complex piece of computer code is intended to do and how it is supposed to relate to the human user. Thinking about software programs *as if* they were agents explains and helps in understanding and designing computational artefacts (Noorman, 2009). Similarly, describing machines as delegates that substitute for human actors helps to draw attention to the role they perform in shaping human actions and morality.

Metaphors, however, are not innocent; they can sometimes even be dangerous. They draw attention to particular similarities between two things, using one that is presumably well understood, to help understand one that is not. However, in thinking metaphorically, we may be directed to think that the two things have

[3] Johnson and Powers (2008) used the metaphor of computer systems as surrogate agents to tease out the possibility of a form of responsibility for computer systems.

more in common than they do. Important and relevant dissimilarities between the compared entities may be pushed to the background by making a particular analogy between the two entities. Moreover, analogies can lead us to believe we understand something when in fact the thing used in the analogy is very poorly understood, e.g., human consciousness. We have to be careful, then, in drawing analogies between relationships in which humans 'act for' other humans and relationships in which artefacts 'act for' humans. For example, income tax preparation software may be thought of as your personal tax accountant (agent), but software and a human accountant are different in important ways.

Moral Autonomy

A third conception of agency involves autonomy. Traditionally, autonomy was thought to be a distinctive feature of humans differentiating them from other kinds of entities. Because they have autonomy we think of humans as having agency. Humans think, choose, decide and then act. Humans act for reasons and their intentional behaviour is outside the ordinary realm of material causality. On the other hand, artefacts do not act for reasons; their behaviour is the result of causality, be it deterministic or non-deterministic.

Human autonomy is what makes morality possible. That is, morality applies to humans and not to animals and machines because humans have autonomy. In moral theory 'ought implies can'. If a being does not have the capacity to freely choose to act (autonomy), then it does not make sense to have a system of moral rules specifying what that being should do. This idea is famously captured in Kant's distinction between things that behave according to natural law and things that behave according to the conception of law. So the autonomy conception of agency is intertwined with a set of ideas about human capacities, action, intentions, and differences between humans and other kinds of beings.

To be sure, this conception of autonomy continues to perplex moral philosophers and many others. Philosophers and ethicists continue to try to explain how it is possible (and whether it is true to say) that humans are free and have consciousness and autonomy. Whether autonomy and consciousness are amenable to reductionist accounts is an issue that will not be taken up here.

Autonomy is also used in other, non-moral contexts to describe artefacts that operate independently from humans. Remember Deborah's refrigerator was thought of as autonomous because it maintained a particular internal temperature without any action on Deborah's part. Similarly, we speak of autonomous vehicles, autonomous systems, autonomous robots, etc. Computer scientists refer to particular programs as autonomous in order to highlight their ability to carry out tasks on behalf of the user and to perform those tasks independently. As a result of machine learning algorithms, for instance, these programs are thought to be more

capable of operating independently in unknown environments than pre-programmed computers systems.

Thus, the autonomy conception of agency seems to include two different conceptions. One has roots in the notion of human moral autonomy and the other refers to the independence of things from immediate control by humans. These two different ideas should not be conflated for human moral autonomy provides a foundation for establishing moral status or moral standing. Because humans are autonomous beings that can choose to act in the world, certain rights can be attributed to them and they can be held responsible for their actions. The other conception of autonomy has little to do with morality or moral standing. It is agency only in the sense that it identifies something as operating independently. Having the capacity to operate independently is not sufficient to justify moral status or standing. For this reason, the autonomy conception of agency has to be used cautiously.

Artefactual Moral Agency

All three conceptions of agency are found in debates about artefactual moral agency, though some authors rely more heavily on one conception or another and some combine or conflate several conceptions. The different conceptions are used to clarify aspects of the role of artefacts in morality. Keeping the three conceptions in mind should facilitate discussion of artefactual moral agency; failure to distinguish them runs the risk, among other things, of overlooking important asymmetries between humans and artefacts. The causal efficacy and acting for conceptions are particularly important in understanding moral consequences but neither has implications for the moral standing of artefacts. The autonomy conception of agency, on the other hand, provides a foundation for moral standing.

The causal efficacy of artefacts is generally the foundation of claims about artefactual moral agency. For example, in Chapter 5 of this volume, Verbeek emphasizes the role of artefacts as mediators. Although Verbeek does not explicitly use the language of causality, his account shows how artefacts affect human experience. Verbeek does not want his account to be interpreted as a claim that in mediating human experience, artefacts entirely determine what happens. This concern belies a causal notion at work in his thinking. In his account, artefacts have an active, but not a final, role in organizing relations between humans and world. In order to better understand this role, he pushes for a distributed or a 'composite' conception of moral agency: agency is not an inherent property of either humans or artefacts; it is the outcome of the interactions between humans and things. He explains that "in their own way – distinct, but not separated – humans and things contribute to moral actions and decisions". 'Contribution' seems here very close to, if not the same as, causal efficacy.

Because of their casual efficacy, artefacts make a moral difference. They make a difference in moral practices, moral outcomes, and even moral notions. Think about the difference in the kind and degree of privacy that individuals have now as so many activities have been configured or reconfigured around computers and information technology. Consider changes in the nature of familial relationships accompanying the use of cell phones, reproductive technologies, and child-rearing devices. Artefacts make a moral difference both for better and worse. A world with aqueducts, bridges, antiseptics, sanitation systems, and bicycles is a world in which humans have more pleasant lives and may live longer. Artefacts facilitate individual moral practices, e.g., playpens help parents keep their children safe and ambulances bring medical treatment quickly to those who need it. Of course, artefacts also work in the other way: landmines help to kill and maim the innocent; electronic devices are used to intrude on personal privacy; and so on. Artefacts affect how we fulfil obligations, keep promises, distribute resources, etc.

So in considering the active role of artefacts in moral actions and practices, we can meaningfully think of artefacts as moral agents by treating agency as causal efficacy in the production of states of affairs or events that have moral consequences. Artefactual moral agency means here simply that artefacts have a role in moral actions and outcomes; they affect or make a difference in moral actions and outcomes.

The 'acting for' notion of agency also grounds a conception of artefactual moral agency, that is, 'acting for' gets at something that is important about the role of artefacts in morality. Artefacts can be said to be moral agents in the sense that (or when) they are delegated tasks that are either constitutive of moral practices or have moral consequences. As Latour suggests, we delegate tasks to artefacts to achieve certain results and when we do so, we effectively treat artefacts as our agents. When artefacts perform delegated tasks that constitute states of affairs with moral features or moral consequences, the artefacts can be thought of as our moral agents. When a hospital machine keeps a person breathing, we might think of the machine as a moral agent; when a cell phone allows parents to keep track of their children, we might think of the cell phone as an assistant in fulfilling parental duties. These artefacts perform delegated tasks that constitute moral practices and have moral significance. In the same way, a landmine might be thought of as an immoral agent in the sense that it has been delegated the task of killing or maiming those who step on it. The landmine acts on behalf of those who have intentionally put it in a particular location.

Remember, however, that the acting for conception of artefactual moral agency is metaphorical and, as mentioned earlier, we have to be careful in using metaphors. Thinking of artefacts as if they are agents that perform tasks on behalf of human actors helps to understand how artefacts can shape moral action. Designers and engineers seem to delegate morality to these artefacts by inscribing their intentions in the design of the artefact in order to affect the user's actions in morally significant ways. However, it is a step too far to claim that humans and artefacts are interchangeable components in moral action. Though they perform some simi-

lar tasks, the traffic light and the police officer directing traffic are not morally the same.

One important difference between humans acting as agents for others and artefacts acting as agents for humans is responsibility. Central to the acting for conception is the idea of delegating tasks; we delegate tasks to humans and to artefacts. Importantly, however, human-to-human delegations generally involve tasks *and* responsibility; human-to-artefact delegations involve tasks but no responsibility. The hospital machine and the landmine perform actions on someone's behalf but neither is (or is considered) responsible (except in a causal sense) for the outcome. We might identify the artefact as the point of failure when we do not get what we expected, but we typically hold the humans who produced or deployed the artefact to blame. Those who design and maintain the artefacts are typically considered responsible both for the accomplishments and failures of artefacts.

There is, thus, a significant and not to be overlooked difference between human-to-artefact and human-to-human delegation relationships. Humans delegate tasks with responsibility to human agents recognizing that the delegate has the capacity to negotiate and re-negotiate with them about appropriate goals and strategies. On the basis of this and the agent's expertise and, often, experience, clients delegate a range of decision-making latitude to the agent and the authority to use it. A literary or press agent is authorized to act on a client's behalf with regard to a particular situation or set of decisions and transactions, e.g., finding a publisher and obtaining a contract for a particular book. The agent may, for example, be authorized to negotiate with others and constrained to develop only the preliminary terms of a contract. Nevertheless, the required actions for these negotiations cannot be fully designated at the beginning or specified in rigid rules or static models of behaviour. The literary agent is authorized to behave as she thinks appropriate in contract negotiations, in order to achieve the desired outcome. She will make judgments in contingent situations drawing on her expertise and background knowledge, and her understanding of her client's wishes.[4] She is responsible for these decisions, and can be called upon to account for and explain her decisions and actions, which may result in praise or blame.

Responsibility is not part of human-to-artefact delegations, that is, when humans delegate tasks to artefacts, they do not delegate responsibility to the artefact. In the delegation of tasks to an artefact, the client defines and redefines, distributes

[4] Because human agents 'acting for' have decision-making latitude and the possibility of renegotiation, trust is an important aspect of human-to-human delegations. Clients must trust that their agents will use their decision-making latitude in accordance with specified constraints. Successful delegation relationships generally build trust, that is, the more an agent acts successfully on a client's behalf, the more the client is likely to trust the agent in the future. Arguably, trust is involved in person-to-artefact delegations. We trust refrigerators to keep our food cold, search engines to bring us relevant results, etc. We speak not just of trustworthy accountants but also of trustworthy (reliable) computing. Of course, what it means to trust a human agent to act on one's behalf and what it means to trust an artefact or a technological system to act on one's behalf are quite different.

and redistributes tasks so that the artefacts behave according to well-defined rules and protocols (Collins and Kusch, 1995). Such delegations may allow variability in artefactual behaviour, but only if humans are indifferent to the variability. For example, the speed with which the hands of the clock move might vary by milliseconds, but the humans who look at the clock do not notice or care. On the other hand, if the artefact behaves in unexpected ways or counter to its delegated task, the artefact is considered flawed; it is not considered irresponsible. A clock can spin its hands clockwise, but when the hands spin backwards, we say it *malfunctions*.

We do not have meaningful practices of blaming clocks or holding them responsible. We might, in fun, say that the clock is behaving badly (stretching the agency metaphor), but we think it is broken. We blame the humans who made the clock. Even were we to have some practice of "blaming clocks" the meaning of saying they are responsible would be blurry at best, and more likely incoherent. When it comes to artefacts, responsibility is traced back to human operators, designers, managers, or even politicians: there is a bug that needs to be fixed, developers or users did not have enough training, or the conditions in which the artefact would be used were not accurately anticipated.

The *acting for* conception of agency, thus, provides another meaningful way to think about artefactual moral agency, but only in a narrow sense. Artefacts have moral agency in the sense that they are delegated tasks that constitute moral practices and have moral consequences. This conception of artefactual moral agency draws attention to certain, previously overlooked, aspects of the role of artefacts, but it does not provide a basis to blur the boundaries between humans and artefacts. The metaphor only goes so far.

Failure to recognize the metaphorical character of accounts describing artefacts as agents acting for humans may lead to conflating this conception of agency with the human autonomy conception of agency. This is especially problematic because the autonomy conception is embedded in a set of ideas that refer to and elucidate the capacity for responsibility. It is tied to what it means to be human. Artefacts do not have the kind of autonomy that has traditionally, and non-reductively, been associated with bearing responsibility for one's actions. Some may argue that it is possible to reduce responsibility to something that applies to artefacts, but to do so would seem to violate something very fundamental about the conception. On the other hand, some may argue that the autonomy conception of agency is *the* only conception of moral agency, i.e., only if artefacts have autonomy can they be considered moral agents. As shown above, this position goes too far and fails to recognize the important role of artefacts in morality.

The autonomy conception of agency is important not just because it grounds morality, but because it confers a particular kind of status. Historically, the autonomy of humans is what distinguished humans from other animals in the chain of being. One reason for distinguishing between humans, animals, and machines is to identify which entities should be accorded rights, especially rights against entities in other categories. Differences in status affect, for example, the right to own

property, to vote, to have freedom of speech, to have privacy and, in daily life to determine one's own actions. Differences in status affect negative and positive rights; they lead to ideas about which entities must refrain from killing or keeping captive which other entities. In the Christian-Judaic moral tradition, entities that have autonomy have a special status. They are accorded rights though they are also assigned responsibility.

In Chapter 8 of this volume, Brey argues against blurring the distinction between humans and artefacts on grounds that doing so diminishes the moral status of humans. He gives two reasons for this. First: "the classical notion of an agent has an important role in our moral image of a human being." Brey argues that when artefacts are called agents, the "special features of human agency are lost" and "the moral image of humans is damaged as a result." Brey's second reason has to do with explaining and accounting for events. He affirms the distinction between explaining events and explaining actions wherein the latter involves reasons and intentions and the former involves causality. He argues that extending the notion of agency to include artefacts will destroy the distinction between actions and events, and eliminate "the special role of actions in our understanding of the world". By embracing a conception of agency that requires autonomy, he draws an ontological line and thereby reserves a special status for human actors, as being the only entities capable of acting. The consequence of this is that humans can still be the centre of morality.

Extending the autonomy conception of agency to artefacts has implications both for diminishing the moral status of humans as well as for increasing the status of artefacts. Implications of the latter kind are perhaps most salient in the discourse on computational artefacts. In the context of computer science, as mentioned before, researchers sometimes use the concept of autonomy to describe how computer systems are capable of performing particular tasks independently. However, when the autonomy conception of agency is used to challenge ontological lines between humans and computers, the conception becomes problematic. Take for instance the discussion that focuses on the idea that human behaviour and cognition can be understood in terms of computational processes. If the autonomy of human beings can be analysed and explained as a series of computations, then it can be formalized and simulated by computers. This discourse is mostly speculative with philosophers and cognitive scientists imagining and speculating that computation may someday produce entities that are not just autonomous but have the capacity for moral autonomy. Some speculate that this will necessitate the granting of rights to these computational entities. Although this will produce only a simulation of human moral autonomy, those who believe deeply in the computational model seem to believe that the equivalence between computational moral autonomy and human moral autonomy would justify the attribution of moral agency to these autonomous computational entities.

Although the suggestion that machines might have moral autonomy seems misconceived, the status of things and humans is not immutable. As mentioned earlier, ANT treats human, natural and artefactual nodes symmetrically in analysis, in

order to discover how they are constructed. In fact ANT does not assume that there are three types of entities *a priori*. Rather the categories, natural, human, and artefactual, are the outcome of negotiations, not a given. From this perspective, the difference in status between humans and artefacts is historically and culturally constituted (Suchman, 1998); humans have been constituted as unique, as the ultimate reference point. This would suggest that status can shift and change. However, such changes would not be without broader consequences to, for instance, social practices in ascribing responsibility or attributing rights.

Using the autonomy conception of agency in relation to artefactual moral agents is problematic in the sense that we are asked to imagine or hypothesize that machines will at some point in their development operate in a way that would justify considering them morally autonomous, ascribing responsibility to them and granting them the status of moral agents. Without knowing how such machines would work, this idea seems to go from using a metaphor to understand certain phenomena, to using it as a basis to attribute status. To be sure, there are and are likely to be in the future, similarities between computational machine autonomy and human autonomy. This is not surprising since humans will build such computational machines to accomplish humanly conceived tasks. However, attributing artefacts a moral status comparable to humans would affect the instrumental status that artefacts now have. Human-artefact relationships have many non-instrumental dimensions, but because of their status, artefacts can be treated merely as means. They are not expected to make judgements based on their own motivations or desires (even if that were possible), nor are they expected to account for their decision-making in the way that humans are. Humans define the boundaries of acceptable behaviour for artefacts, and allow them to operate within these constraints. By moving from metaphor to status these complementary statuses would be compromised.

Conclusion

Our analysis suggests, then, that attributions of moral agency to artefacts make sense when they refer to the causal efficacy of artefacts and when they refer to the tasks that have been delegated to artefacts by humans. There may well be other intelligible accounts of artefactual moral agency, but we have argued that using autonomy as the basis for artefactual *moral* agency is problematic. Attempts to extend moral autonomy to artefacts seem to move from a metaphor to a claim of moral status, that is, they claim humans and machines are analogous and, then, on the basis of the analogy attribute to artefacts the status (or potential to have the status) associated with moral autonomy.

However, there are good reasons for keeping the status of humans and artefacts different, that is, for keeping humans and artefacts in different categories with regard to agency. For instance, in separated categories they can be treated as com-

plementary rather than equivalent. More importantly, in order to maintain human responsibility for the development and deployment of artefacts, an anthropocentric moral perspective is essential (Johnson and Miller, 2009). Whether we justify attributions of moral autonomy and responsibility on utilitarian or non-utilitarian grounds, attributions of moral responsibility (the expectation that one will be held to account) have the effect of shaping human behaviour. And if one wants to shape the behaviour of artefacts, then we ought to hold onto practices that hold humans responsible for their design and deployment.

To be sure, the three conceptions distinguished above are intertwined and cannot be strictly separated, which makes the question of responsibility enormously complex. The causal efficacy of artefacts – the availability and behaviour of artefacts – affects the moral autonomy of human agents. What a human can and cannot do is often a function of the artefacts – the built world – constituting the situation. Attributions of responsibility to humans are, then, intertwined with the artefacts with which they act. A person who intentionally launches a computer virus could not have produced, and could not have achieved, the resulting effects were it not for the non-human components that constitute the computers used and the network of the Internet. Artefacts facilitate, persuade, discourage and sometimes prevent humans from taking actions or making particular decisions.

Moral agency could be extended to the whole sociotechnical system. We might say that it is not the human or the gun that performed the moral action of killing someone; the actor at issue is the combination of the two. Or we might say that it is the entire sociotechnical system, i.e., the gun, the human, the arms manufacturer, the gun seller, the policymakers that, all together, allowed the gun to be fired. The result of such distributions, however, could be that responsibility is everywhere and nowhere.

The challenge is, then, how to handle distributed responsibility. Here there is something to be gained from holding humans morally responsible for the artefacts they make and for setting the boundaries within which artefacts are allowed to operate. This may be seen as a conservative position, but it is not conservative when taken as an anchor in the endeavour to address distributed responsibility and to frame the challenge of artefact design and use.

References

Barad, K. (1996). Meeting the universe halfway: Realism and social constructivism without contradiction. In J. H. Nelson & J. Nelson (Eds.), *Feminism, science and the philosophy of science* (pp. 161-194). Dordrecht: Kluwer Academic Publishers.

Collins, H. and M. Kusch (1998). The shape of actions. What humans and machines can do. Cambridge, Massachusetts: The MIT Press.

Griffin, J. (2008). *On human right*s. Oxford: Oxford University Press.

Heath, J. (2009) The uses and abuses of agency theory. Business Ethics Quarterly, 19(4), 497-528, 32 p. October 2009.

Himma, K. E. (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.

Johnson, D. and K. Miller (2008). Un-making artificial moral agents. Ethics and Information Technology, 10(2-3), 123-133.

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In *Shaping technology/building society. Studies in sociotechnical change*, edited by W. E. Bijker & J. Law, Massachusetts: The MIT Press, 225–258.

Law, J. and J. Hassard (1999). Actor network theory and after. Oxford, UK/Malden, MA: Blackwell Publishers/The Sociological Review.

Lee, N. and S. Brown (1994). Otherness and the actor network. American Behavioral Scientists 37(6), 772-790.

Liao, S. M. (2010). Agency and human rights. Journal of Applied Philosophy, 27(1), 15-25.

Napoli, C. (2007). Software agent negotiation for service composition. In *Agent and multi-agent systems. Technologies and applications. Lecture notes in computer science*, Vol. 4496, 456-465.

Noorman, M. (2009). Mind the gap a critique of human/technology analogies in artificial agent discourse. Universitaire Pers Maastricht.

Rahwan, I., L. Sonenberg, N. Jennings, and P. McBurney (2007). Stratum: A methodology for designing heuristic agent negotiation strategies. Applied Artificial Intelligence, 21(6), 489-527.

Suchman, L. (1998). Human/machine reconsidered. Cognitive Studies 5(1), 5-13.

Suchman, L. (2001). Human/machine reconsidered. Published by the Department of Sociology, Lancaster University at: http://www.comp.lancs.ac.uk/sociology/soc040ls.html.