Sci Eng Ethics (2017) 23:431–448
DOI 10.1007/s11948-016-9802-1

ORIGINAL PAPER

Distributed Cognition and Distributed Morality: Agency, Artifacts and Systems

Richard Heersmink[1]

Received: 24 February 2016 / Accepted: 10 June 2016 / Published online: 5 July 2016

Abstract There are various philosophical approaches and theories describing the intimate relation people have to artifacts. In this paper, I explore the relation between two such theories, namely distributed cognition and distributed morality theory. I point out a number of similarities and differences in these views regarding the ontological status they attribute to artifacts and the larger systems they are part of. Having evaluated and compared these views, I continue by focussing on the way cognitive artifacts are used in moral practice. I specifically conceptualise how such artifacts (a) scaffold and extend moral reasoning and decision-making processes, (b) have a certain moral status which is contingent on their cognitive status, and (c) whether responsibility can be attributed to distributed systems. This paper is primarily written for those interested in the intersection of cognitive and moral theory as it relates to artifacts, but also for those independently interested in philosophical debates in extended and distributed cognition and ethics of (cognitive) technology.

Keywords Moral status of artifacts · Moral agency · Material agency · Systems agency · Neuroethics · Responsibility · Distributed moral cognition

Introduction

It has been argued that our cognitive and moral capacities are, under certain conditions, distributed across human agents and artifacts. Distributed cognition theory (Hutchins 1995) and the closely related theory of extended cognition (Clark 1997, 2008) developed in cognitive science (See "Extended and Distributed Cognition" section). Distributed morality theory developed in ethics of technology (Magnani and Bardone 2008; Floridi 2013; Verbeek 2011). These theories in

✉ Richard Heersmink
richard.heersmink@gmail.com

[1] Department of Philosophy, Macquarie University, Sydney, Australia

seemingly quite distinct subfields actually have a lot in common regarding their approaches and the ontological status they attribute to artifacts and the larger systems of which they are part. Both approaches emphasize the importance of artifacts for better understanding human capacities and both take a systems view. In comparing and evaluating these views, I argue that some artifacts, depending on the way they are used, have cognitive and moral status, but lack cognitive and moral agency. Whilst artifacts "do" things for and to their users (i.e., they are goal-realisers and have transformative effects on cognition and moral reasoning), using the term "agency" to describe the things artifacts "do" and the effects they have is misleading and inconsistent with traditional notions of agency. I shall argue that extended cognitive systems can be said to have agency only when the artifact is fully transparent and densely integrated into the cognitive processes of its user, whereas distributed cognitive systems without central control lack agency (See "Distributed Morality" section). Having evaluated and compared these views, I then conceptualise how cognitive artifacts are used in moral practice. I specifically focus on the way such artifacts scaffold and extend moral reasoning, have a certain moral status which is contingent on their cognitive status, and whether responsibility can be attributed to distributed systems (See "Moral Practice, Artifacts and Responsibility" section).

Extended and Distributed Cognition

Anthropologist Ed Hutchins (1995), philosopher Andy Clark (1997, 2008), and various others (Rowlands 1999; Kirsh 2006; Sutton 2006, 2010; Menary 2007; Wheeler 2010) developed a view on human cognition, arguing that cognitive states and processes are, in some cases, distributed across humans and artifacts. When that happens, human agents together with technological artifacts form an integrated system that performs information-processing tasks. On this view, "a cognitive process is delimited by the functional relationships among the elements that participate in it, rather than by the spatial colocation of the elements" (Hollan et al. 2000, p. 176). Thinking or cognizing is thus not something that takes place exclusively in human brains, but can be distributed across brains, bodies, and environment. Hutchins' main examples of distributed cognitive systems include a team of seafarers interacting with artifacts and instruments on a navy ship, airline pilots interacting with cockpit equipment, and people interacting with computer systems. Clark's main examples of extended cognitive systems include a man with Alzheimer's disease using a notebook to supplement his poor biological memory, making a difficult calculation with pen and paper, writing an academic paper with a word-processor, and sketching preliminary structures during a design or artistic process.

On the basis of these examples, one can infer that Clark's extended cognition theory focusses on single agents interacting with artifacts, whereas Hutchins distributed cognition theory typically (though not exclusively) on larger systems with more than one agent interacting with artifacts. In such wider cognitive systems, there are thus one or more individuals interacting and coupling with cognitive artifacts. These artifacts provide their users with information that is necessary for performing a cognitive task, e.g., navigating a ship, landing an airplane, writing a document on a computer, or making a calculation. Sometimes such artifacts provide their users with information that is fixed or static (e.g., a checklist pilots use before taking off) and sometimes it is dynamic and changes during a task (e.g., writing a document on a computer). In the latter case, it allows agents to manipulate and process the external information in a way that is difficult to do in their brains, in that way *complementing* information-storage and processing properties of human brains (Sutton 2010).

The Cognitive Status of Artifacts and Systems

Whilst there are epistemological and methodological differences between extended and distributed cognition theory (Hutchins 2014; Smart et al. 2016), they both share an ontological commitment to the notion of artifacts as co-constitutive of a larger cognitive system. So, on these views, extended or distributed cognitive systems have components that interact, transform each other, and are integrated into a wider cognitive system. But when exactly does that happen? Clark and Chalmers (1998; see also Clark 1997, p. 217) introduce a number of criteria to distinguish between artifacts that are part of an extended cognitive system and those that merely aid and scaffold one’s cognition. They argue that what is central for external information to be co-constitutive of a cognitive state or process is a high degree of trust, reliance, and accessibility, and we most have endorsed it at some point in the past. Thus when Otto, a man with Alzheimer’s disease and a poor biological memory, uses a notebook as an external memory device, the information in the notebook is reliable, trustworthy, easily accessible, and has been endorsed at some point in the past. For these reasons, Clark and Chalmers argue, the notebook and the information in it are co-constitutive of Otto’s cognitive system.

While these criteria provide a helpful starting point for thinking about when an artifact or other resource is part of an extended cognitive system, a number of theorists have questioned some of these criteria (e.g., Michaelian 2012; Ludwig 2015) or added other criteria (Menary 2010; Sterelny 2010). Clark and Chalmers’ criteria invite us to think of extended cognitive systems as a black or white phenomenon. Cognitive systems are either genuinely extended, or they are not. When a system satisfies these criteria, it is seen as a genuine extended system and when it does not sufficiently satisfy these criteria, it is seen as an embedded or scaffolded cognitive system.

Given the complexity and multidimensional nature of the way we interact with cognitive artifacts, I suggest, we should not take a threshold view on membership of extended cognitive systems. It is better to conceive of system membership in terms of the degree of cognitive integration of humans and artifacts. This integration, in turn, depends on a number of dimensions (Sutton 2006, 2010; Wilson and Clark 2009; Menary 2010; Sterelny 2010). The dimensions of this spectrum include the kind and intensity of information flow between agent and scaffold, the accessibility of the scaffold, the durability of the coupling between agent and scaffold, the amount of trust a user puts into the information the scaffold provides, the degree of transparency-in-use, the ease with which the information can be interpreted, the amount of personalization, and the amount of cognitive transformation (Heersmink 2015). Cognitive artifacts that rank high on these dimensions are integrated deeper than those that rank low on these dimensions. When artifacts are integrated deeply, they have cognitive status, that is, they are part of an extended cognitive system. When artifacts are integrated shallowly, they do not have cognitive status. Because cognitive integration is a complex and multidimensional phenomenon, it is difficult to demarcate clearly between artifacts with cognitive status and those that lack this status. There is thus no clear tipping point. The paradigm cases are clear, but there is a grey area in between.

It is important to note that extended and distributed cognition theory do not claim that artifacts are cognitive in themselves: only when used and integrated in the right sort of way do artifacts become part of a wider system and in that way obtain cognitive status. Thus, only when such artifacts are actually being used, i.e., when their informational properties realise their cognitive functions, do they obtain cognitive status. In their inactive and dormant state, so to speak, they are mere objects with the potential to obtain cognitive status. So Otto's notebook in itself does not belief anything, only Otto does. As Clark points out:

> The appeal to coupling is not intended to make any external object cognitive (insofar as this notion is even intelligible). Rather, it is intended to make some object, which in and of itself is not usefully (perhaps not even intelligibly) thought of as either cognitive or noncognitive, into a proper part of some cognitive routine. It is intended, that is to say, to ensure that the putative part is poised to play the kind of role that itself ensures its status as part of the agent's cognitive routines (Clark 2008, p. 87).

The point of coupling and integration is that an artifact or other external resource becomes "part of the physical substrate of a cognitive system" (Clark 2008, p. 88), not that it becomes cognitive in itself. The wider system is thus cognitive, but not the artifact.

The Cognitive Agency of Artifacts and Systems

Under particular circumstances, some artifacts thus have cognitive status because the artifact and its user are densely integrated, in that way forming a wider cognitive system. Does their cognitive status imply they have cognitive agency? It seems to me that artifacts do not have cognitive agency because they are not cognitive agents, that is, they do not have the capacity to cognize. In order for something to have cognitive agency, it must have the capacity to initiate thoughts and mental states such as beliefs, desires, or intentions. Artifacts come into existence through human intentions and agency, but do not have themselves intentions (or other mental states) and thus lack agency, cognitive or otherwise. This is not to say that artifacts are mute or inert objects, they are not. Artifacts "do" things for and to their users: they are active and have transformative effects on the cognitive skills of their users. But using the term "agency" to describe the things artifacts "do" and the effects they have is misleading and conceptually confusing. Attributing cognitive agency to artifacts seems to obscure important differences between real cognizers (i.e., humans and certain animals) and artifacts.

But what about the larger extended or distributed system? Can we attribute cognitive agency to such systems as a whole? In his early work, Clark (1997) speculates about the possibility of extended agency but ultimately remains agnostic. “In sum, I am content to let the notions of self and agency fall where they will” (1997, p. 218). Ronald Giere, by contrast, has stronger views on systems agency. In relation to distributed cognitive systems in scientific practice, he argues that both artifacts and the larger systems lack cognitive agency. “It is the humans, and only the humans, that provide intentional, cognitive agency to scientific distributed cognitive systems. We need not extend our notions of cognitive agency to include other components of these distributed cognitive systems” (Giere 2004, pp. 772–773). He later argues that: “We should regard the human components of distributed cognitive systems as the only sources of agency within such systems. In particular, we should not extend notions of agency to such systems as a whole” (Giere 2006, p. 710). I agree with Giere that we should not attribute cognitive agency to artifacts, but I partly disagree with Giere on systems agency.

What is an agent and which entities have agency? Briefly, an agent has the capacity to generate and realise its intentions. We typically do not say that the brain—or, more specifically, the prefrontal cortex—is the only relevant component for being an agent because it generates our intentions. The entire embodied organism is needed to generate intentions and act in the world. We do not think of the human body as a mere instrument of the intentional mind, as that would result into Cartesian dualism. Rather, body and mind are a single, integrated unit. This anti-dualism can (and should) be pushed further. A blind man using a cane or indeed Otto using his notebook, for example, do not see the cane and notebook as mere objects in their lifeworld, but as part of the apparatus with which they encounter and act in the world. These objects are procedurally and cognitively transparent in use, that is, the blind man and Otto do not need to think about *how* to use these objects to achieve their goals. They have used these objects so often that they became a transparent part of their behavioural routines in a similar way as one’s body is transparent (Clark 2008). The embodied brain, evolved to be a pragmatic and opportunistic system, does not care whether these objects are artifactual rather than biological. What matters is whether they can be used to realise intentions. An embodied-organism-plus-such-transparent-tools should be seen as the system that realises intentions, not just the embodied organism.

In an analysis of the relations between extended and distributed cognition, Hutchins (2014) points out that extended cognition is a subset of cognitive events that involve interaction between internal (biological) and external (technological) resources. By contrast, distributed cognition is much broader as it is a view on all of cognition. On Hutchins view, the question is not whether or when cognition is distributed. “Rather, the interesting questions concern the elements of the cognitive system, the relations among the elements, and how cognitive processes arise from interactions among those elements” (2014, p. 36). Additionally, extended cognition is agent-centered in that it conceptualises a human agent as the center, controller and assembler of extended cognitive systems. Distributed cognition theory does not assume that humans are necessarily the center of distributed systems. “Centers and boundaries are features determined by the relative density of information flow across a system” (Hutchins 2014, p. 37).

One possible way to look at systems agency is to say that an extended cognitive system *is* an agent because it is a system capable of generating intentions and realising them by means of a densely integrated and transparent artifact. By contrast, at least some distributed cognitive systems are *not* agents because control is much more decentralised. Hutchins points out that “Some systems have a clear center while other systems have multiple centers or no center at all” (2014, p. 37). Giere, for example, conceptualises particular scientific laboratories as distributed cognitive systems, but these systems do not always have a human as a central controller whose intentions are being realised. Rather, they are large-scale, bottom-up systems with decentralised control. So, Otto and his notebook are an extended agent because when Otto uses his notebook, he uses information *he* made *himself* to realise *his* goals. In contrast, when experiments are conducted in a scientific laboratory, the various researchers and artifacts and instruments do not constitute an extended agent because there is no central agent whose intentions are being realised. Thus, I agree with Giere that at least some (or perhaps even most) distributed cognitive systems are not agents, but disagree with Giere by arguing that extended cognitive systems are agents and thus have agency (see Malafouris 2008; Knappett and Malafouris 2008 for further discussion on material agency).

Distributed Morality

Floridi on Distributed Morality

The idea that morality or moral agency is distributed across humans and technology is developed by at least two philosophers, namely Luciano Floridi and Peter-Paul Verbeek. Floridi (2013; Floridi and Sanders 2004) uses the phrase “distributed morality” to refer “only to cases of moral actions that are the result of otherwise morally-neutral or morally-negligible interactions among agents constituting a multiagent system, which might be human, artificial, or hybrid” (2013, p. 729). By using the notion of distributed knowledge in epistemic logic, Floridi helps to clarify and motivate his notion of distributed morality. Person A knows that either the car is in the garage or Jill got it. Person B knows that the car is not in the garage. Neither person A nor B knows that Jill got the car, but the supra-agent, person-A-plus-B, knows that Jill got the car. So, new knowledge is created, only when these two epistemic states are integrated in the right sort of way. In this sense, the building blocks of new knowledge are distributed. According to Floridi, morality is distributed in a similar way. The idea is that a morally-relevant action can result from many small morally-neutral or morally-negligible actions. So, when many morally-neutral or negligible actions are accumulated, a morally-relevant action can result.

Let's further unpack the above definition and look at some examples. A multiagent system may consist of two or more human agents, two or more artificial agents, or some combination of human and artificial agents. Floridi's examples are a corporation, an autonomous network of drones, or a human navigating with a GPS device. Many interactions between agents, either human or artificial, within a multiagent system often turn out to be morally neutral or negligible. However, in some cases, a chain or set of interactions in a multiagent system may overcome a threshold and have morally significant consequences, either negative or positive. Floridi adopts a non-anthropocentric approach to morality, arguing that "we need to evaluate actions not from a sender but rather from a receiver perspective: actions are assessed on the basis of their impact on the well-being of the environment at large and its inhabitants specifically" (2013, p. 731).

One specific example of distributed morality Floridi mentions is a customer loyalty scheme of a bank. The bank offers credit cards which are linked to specific charities such as Oxfam. When customers open a credit card account, the bank automatically donates £15 to charity. If the account is used within 6 months, a further £2.5 is donated. And for every £100 spent with the card, another £0.25 is donated. So this multiagent system comprises customers, bank employees, software agents, and other infrastructural components, interacting in such a way that one single interaction does not make a significant moral difference, but many interactions do. I take it that the point is not so much to determine the threshold of when a number of interactions become morally relevant, but to make clear that the accumulation of many morally negligible interactions ultimately result in a positive moral outcome. In this example, morality is distributed across the entire multiagent system, including the human and software agents. Morality can thus be seen as a property of a multiagent system, not exclusively of individual humans.

Verbeek on Distributed Moral Agency

By drawing on actor-network theory, Verbeek (2011) develops the notion of "distributed moral agency". He specifically focusses on the moral aspects of obstetric ultrasound imaging, an imaging technology used to visualize a foetus in the mother's womb, which is usually done at week twelve and twenty of pregnancy. This technology does not just provide a neutral and transparent window to the womb, but has several morally important aspects. First, it creates an enlarged image of the foetus, making it look somewhat independent and isolated from the mother's body. It also reveals the foetus' gender, allowing parents to call the unborn by its name. Verbeek argues that these aspects give the foetus a kind of personhood. Second, and more important, it shows possible genetic conditions such as Down syndrome or heart disease. This translates the unborn foetus into a possible patient about which moral decisions need to be made. So obstetric ultrasound imaging mediates and transforms the relationship parents have to their foetus, providing them with information on the health of their foetus that is important for moral decision-making regarding pregnancy and abortion. For these reasons, moral agency should be seen as distributed across humans and technology. In Verbeek's words:

> Ultrasound imaging actively contributes to the coming about of moral actions and the considerations behind moral actions. This example therefore shows that moral agency should not be seen as an exclusively human property; it is distributed among human beings and nonhuman entities (2011, p. 38).

In a later chapter he points out:

> Moral mediation always involves an intricate relation between humans and nonhumans, and the "mediated agency" that results from this relation therefore always has a hybrid rather than a "purely human" character. When technologies are used, moral decisions are not made autonomously by humans, nor are persons forced by technologies to make specific decisions. Rather, moral agency is distributed among humans and nonhumans; moral actions and decisions are the product of human-technology associations (2011, p. 53).

Verbeek here argues that moral agency is distributed across humans and technology, they cannot be understood in isolation from each other because they are integrally connected. So, one way to understand the moral relevance of technological artifacts is to say that artifacts are co-constitutive of moral agency.

Discussion

Both Floridi and Verbeek argue that moral actions, either positive or negative, can be the result of interactions between humans and technology, giving artifacts a much more prominent role in ethical theory than most philosophers have. They both develop a non-anthropocentric systems approach to morality. Floridi focuses on large-scale "multiagent systems", whereas Verbeek focuses on small-scale "human–technology associations". But both attribute morality or moral agency to systems comprising of humans and technological artifacts. On their views, moral agency is thus a system property and not found exclusively in human agents. Does this mean that the artifacts and software programs involved in the process have moral agency? Neither of them attribute moral agency to the artifactual components of the larger system. It is not inconsistent to say that the human-artifact system has moral agency without saying that its artifactual components have moral agency. Systems often have different properties than their components. The difference between Floridi and Verbeek's approach roughly mirrors the difference between distributed and extended cognition, in that Floridi and distributed cognition theory focus on large-scale systems without central controllers, whereas Verbeek and extended cognition theory focus on small-scale systems in which agents interact with and control an informational artifact. In Floridi's example, the technology seems semi-autonomous: the software and computer systems automatically do what they are designed to do. Presumably, the money is automatically transferred to Oxfam, implying that technology is a mere cog in a larger socio-technical system that realises positive moral outcomes. There seems to be no central controller in this system: it is therefore difficult to see it as an extended agency whose intentions are being realised.

It is noteworthy that Floridi motivates and explains his view in part by appealing to the notion of distributed knowledge in epistemic logic. The example he gives of a supra-agent (person-A-plus-B) knowing more than its constituent members (persons A and B) is basically a case of socially distributed cognition. Compare the following real-world example from Harris et al. (2010) of socially distributed remembering. In this example, a long married couple recalls the name of the show they saw on their honeymoon more than 40 years ago. Neither of the constituent members knows the answer straight away, but by interacting in the right sort of way, they jointly construct the answer.

F: And we went to two shows, can you remember what they were called?
M: We did. One was a musical, or were they both? I don't… no… one
F: John Hanson was in it
M: *Desert Song*
F: *Desert Song*, that's it, I couldn't remember what it was called, but yes, I knew John Hanson was in it
M: Yes

This example of collaborative remembering is, in terms of distribuends, very similar to Floridi's example of distributed knowledge. In both examples, the outcomes (i.e., "Jill got the car" and "Desert Song") are achieved by integrating two epistemic states which are more than the sum of their parts, but in Floridi's examples of distributed morality, the outcome is exactly the sum of its parts (i.e. the outcome is the sum of all donations). Every time someone opens a credit card account or uses it, a certain amount of money is donated to charity, which is cumulative, not emergent. So whilst I am sympathetic to Floridi's example of distributed knowledge, I am not sure whether it supports or is analogues to his view of distributed morality. More generally, we can ask whether extended or distributed cognition can be used to motivate a notion of distributed morality. I think it can, but it will result in a different notion of distributed morality than the one Floridi is advocating. Below I argue that moral cognition can be extended and distributed by interacting with artifacts or other people, in that way resulting in cognitive systems that make moral decisions and morally-relevant changes in the world. Such systems are distributed moral systems.

In Verbeek's example, the technology provides information that parents use to make important moral decisions about their unborn child, but the parents make the decision, not the technology. The technology might invite certain actions, but ultimately the parents' intentions result in a particular moral outcome. Of course, the decision whether or not to abort would have been different if the imaging technology would not be available and so the technology deeply influences moral decision-making. But when is an artifact part of a distributed moral system? Are there criteria for thinking about when an artifact is part of a distributed moral system? Neither Floridi nor Verbeek explicitly provide such criteria. But on the basis of their claims and examples, one may infer that the artifact needs to enable or at least influence the moral outcome of the larger socio-technical system. Without software agents the money would not be transferred to charity and without ultrasound imaging there would be no information to make decisions about the foetus. Remove the artifact from the system and there will be a different moral outcome. As Verbeek writes: “If one the two were missing, this type of agency could not exist” (2014, p. 80). But is this sufficient for the artifact to be co-constitutive of a distributed moral system? I now suggest it is not.

Above I wrote that a distinction is made between embedded and extended cognitive systems. In an embedded cognitive system, an artifact aids and influences cognition but is not part of the cognitive system, whereas in an extended cognitive system, an artifact is part of a cognitive system. Whether an artifact is part of a cognitive system, it has been suggested by a number of philosophers (Sutton 2006; Menary 2007, 2010; Sterelny 2010; Heersmink 2015), depends on how it is integrated into the cognitive processes of its user. Thus, using traffic signs to navigate an unfamiliar city may aid and influence cognition, but those signs are not part of an extended cognitive system because they are not deeply integrated into the cognitive processes of their users. By contrast, personalized maps on one’s mobile and context-aware computing device might be co-constitutive of cognition if they are integrated in the cognitive processes of its user in the right sort of way. The degree of integration depends on how a system ranks on the dimensions briefly outlined in “The Cognitive Status of Artifacts and Systems” section. Personalised maps rank higher on the dimensions of information flow, accessibility, durability, personalization, and transformation. For these reasons, such maps are integrated more deeply into the cognitive processes that govern navigation than traffic signs are and are thus better candidates for extended cognitive systems.

The point here is that there is a multidimensional spectrum between embedded and extended cognitive systems. Applying this kind of view to distributed morality, it seems that Verbeek’s example is perhaps better seen as an embedded moral system, rather than an extended or distributed one. Moral decisions, in Verbeek’s example, depend on the information that imaging technology provides, but that information does not seem to be deeply integrated into moral reasoning in the same way that cognitive artifacts are sometimes deeply integrated into cognitive processes. There is no informational reciprocity between technology and agent, so the information that the technology provides is merely input to the brainbound (but embedded) cognitive system of the users.

In an evaluation of Verbeek’s position, Philip Brey makes the following point. “While I agree with Verbeek that human agency is often influenced by artifacts, and I am even willing to agree that agency can be attributed to human-artifact assemblies, it does not follow that artifacts therefore have some form of agency…” (2014, p. 135). Like the majority of philosophers, Brey denies that artifacts have agency: he does, however, accept the notion of systems agency. I think systems agency occurs only when an agent uses a densely integrated and transparent artifact to realise its intentions. The relationship the parents have to the imaging technology does not seem to satisfy these criteria. However, for the technician or gynaecologist who is directly interacting with the technology, it could potentially be such a transparent and integrated tool, depending on the way it is used.

Finally, Verbeek wants to argue that all moral agency is the result of human–technology associations. “Morality is a hybrid affair; it cannot be located exclusively in things, but not in humans either” (2014, p. 80). In parallel to the notion of cognitive bloat (which is a *reductio ad absurdum*, claiming that cognition unrealistically extends into too many artifacts), this might be called "moral bloat". *Even* if we grant that moral agency can be distributed, there is no reason to think that *all* moral agency is distributed. Many moral decisions and actions are made without interacting with technology. To avoid moral bloat, i.e., that moral agency is unrealistically distributed across too many artifacts, distributed morality can learn from extended cognition to develop explicit criteria for deciding when an artifact is part of a distributed moral system. Conversely, one of the lessons extended and distributed cognition theory might learn from distributed morality is to include moral dimensions of artifacts into their conceptualisation of wider cognitive systems, which I do in the next section.

Moral Practice, Artifacts and Responsibility

Having evaluated and compared extended and distributed cognition with distributed morality theory, I now continue by outlining how cognitive artifacts are used in moral practice. I single out three moral dimensions of such artifacts, namely their functional role in moral reasoning, their moral status being contingent on their cognitive status, and whether moral responsibility can be attributed to distributed systems.

Embedded and Extended Moral Cognition

The paradigm examples in the extended and distributed cognition literature concern cognitive processes such as navigating (Hutchins 1995), mental imagery (Clark and Chalmers 1998), remembering (Rowlands 1999; Michaelian and Sutton 2013), and calculating (Clark and Chalmers 1998). Moral cognition, however, seems to be absent. Furthermore, these paradigm examples typically (but not necessarily) exhibit reciprocal information flow or what Clark (1997) calls "continues reciprocal causation". Partly due to this reciprocity, the artifact is integrated deeply into cognitive processes of its user. Examples are making a complex calculation with pen and paper, sketching on paper or canvas, and writing an (academic) article. In these cases, there are various cycles of informational offloading and interpretation in which each cycle depends on the outcome of the previous one.

How might moral cognition be extended in a similar way? Dealing with a moral dilemma and subsequently making a moral decision is often a complex cognitive process. As April Martin, Zhanna Bagdasarov and Shane Connelly recently point out in this journal:

> In order to successfully handle such dilemmas it is necessary to represent many pieces of information in mind at once. Additionally, the decision-maker is tasked with determining which factors are important, how they relate to the dilemma and to each other, and sometimes even how various interactions of these factors might come into play (2015, p. 276).

Their empirical research shows that in making moral decisions working memory capacity has a “unique variance above and beyond ethics education, exposure to ethical issues, and intelligence” (2015, p. 282). Our working memory thus plays an important role in moral cognition, but is limited in dealing with all the relevant information. To overcome these limits and to make sense of the moral situation, some people write a list of pros and cons when making an important moral decision, in that way better overseeing the situation and consequences of an action. When two or more moral actions involve various possible positive and negative outcomes, it is difficult for the human brain to oversee, compare, and weigh all the options. Offloading information onto an artifact helps to overcome these limits and allows one to compare various options. Externalizing information can thus improve moral decision-making by helping an agent to make sense of a moral situation (see also Johnson et al. 2013). In this way, making lists can extend an agent’s moral reasoning processes.

Ethical matrices can similarly extend an agent’s moral reasoning. Ethical matrices are cognitive tools to analyse ethical problems and aid decision-making. Mepham (2000), for example, developed an ethical matrix to help people make decisions about biotechnology, but is also applicable for other technologies. The matrix has a tabular format: it has a column for relevant interest groups and stakeholders and three columns for key values such as well-being, autonomy, and justice (broadly construed). To better understand the moral situation and how a technology influences each stakeholder, the matrix allows its user to briefly outline how each value plays out for each stakeholder. This allow the users of the matrix to include different viewpoints, in that way typically making a better decision. Such a matrix is “at its simplest level a checklist of concerns, structured around established ethical theory. However, it can also be used as a means of promoting structured discussion” (Kaiser et al. 2007, p. 71). So, depending on the way it is used, its cognitive functions may range from merely checking how each stakeholder is affected by a technology to facilitating sustained moral debate with a group of people.

Moral reasoning might also be merely scaffolded or embedded by using cognitive artifacts. Certain professions, such as engineering, have moral codes of conduct. Such codes include guidelines for engineers’ responsibility to the public, clients, employers, colleagues, and their selves. Often such codes have both general principles and more specific guidelines. For example, the Code of Ethics and Professional Conduct of the Association for Computing Machinery[1] includes general principles such as: “Software engineers shall be fair to and supportive of their colleagues.” It also contains more specific guidelines such as: “In particular, software engineers shall, as appropriate, credit fully the work of others and refrain from taking undue credit.” There will be cases in which an engineer is confronted with a moral problem that he or she is unable to solve or deal with on one’s own (Davies 1991). In such cases, an engineer can consult the code of ethics whose specific principles might aid one in making better grounded moral decisions, in that way scaffolding an engineer’s moral reasoning and decision-making processes. A

[1] http://www.acm.org/about/se-code.

code of ethics merely scaffolds and not extends cognition, because the information it contains is made by someone else. The information flow is thus only one-way and not reciprocal. For this reason, the cognitive integration between agent and external information is rather shallow.

Moreover, it is not just artifacts that can scaffold or extend one's moral reasoning. Neil Levy suggests that moral cognition can also be socially distributed. "Moral thought, too, should be thought of as a community-wide enterprise" (Levy 2007b, p. 309). I agree with Levy, but he focusses on how moral knowledge disseminates through society as a whole, whereas I think smaller groups like dyads, families, a group of friends, a team of colleagues, a class of students, juries, and so on, are more tractable cases of socially distributed moral cognition. Such small social groups often discuss moral problems and make moral decisions as a group. For instance, a team of engineers may face a moral dilemma when designing a certain product or system, say, releasing an air-control system with a minor bug, potentially causing safety issues, or not releasing it, potentially causing many people to get fired. In such moral debates, there often (though not necessarily) is a high degree of informational exchanges and reciprocity between group members, in that way forming a socially distributed cognitive system that makes morally-relevant decisions.

The Moral Status of Cognitive Artifacts

As argued above, attributing moral agency to artifacts is controversial and many philosophers have argued against it (Johnson 2006; Himma 2009; Peterson and Spahn 2011). However, denying that artifacts have moral agency does not mean that they have no moral relevance, they do. One aspect of the moral relevance of artifacts is that they have a certain moral status. Clark and Chalmers point out that, on the extended mind view, "in some cases interfering with someone's environment will have the same moral significance as interfering with their person" (1998, p. 18). On the basis of these considerations, Levy (2007b) argues that internal and external mental states are ethically on a par (compare DeMarco and Ford 2014). "If we worry, say, that enhancing the brain pharmacologically is (for whatever reason) wrong, or that transforming it using, say, magnetic stimulation or surgery risks authenticity, then we should worry equally about analogous interventions into the extended mind" (2007b, p. 61). Traditionally, neuroethics (in general) and the cognitive enhancement debate (in particular) are not so much interested in ethical aspects of external artifacts. Its main focus is on psychopharmaceuticals and to a lesser extent on emerging technologies like brain–computer interfaces, transcranial magnetic stimulation, and neuroprosthetics. Levy (2007a, b) points out that if the extended mind thesis is correct, then neuroethics should expand its scope to include moral reflection on environmental objects and structures. Note that this insight can potentially create an interesting link between neuroethics and ethics of technology. The point here is that, if mind and cognition are, in some cases, partly constituted by external information, we ought to treat that information in a similar way as we treat mental states instantiated purely by the brain.

More concretely, if information in Otto's notebook is constitutive of his cognitive system, then tampering with this information is the same as tampering with information stored in his biological memory. For this reason, Johnny Søraker claims that "the case with Otto's notebook suggests that information and information technology can have moral status, but only if they are constitutive and irreplaceable in a strong sense" (2007, p. 14). The moral status of artifacts is thus contingent on their cognitive status. So, artifacts that are part of extended and distributed cognitive systems have moral status, whereas artifacts that are part of embedded cognitive systems lack this status. In addition to traditional legal and moral aspects of artifact ownership, it implies that we ought not to interfere with people's wider minds. In general, the more we depend on a cognitive artifact for our everyday functioning, the deeper it is integrated into our cognitive system (Heersmink 2016). Stealing or altering Otto's notebook should thus not only be seen as illegal, but also as morally problematic because he deeply depends on it for his everyday cognitive functioning in the same way healthy subjects depend on information in their biological memory for their everyday functioning.

Distributed Responsibility

Anthropologist F. Allan Hanson (2009) argues that what he calls "extended agencies" can be held responsible for their actions. Hanson's notion of extended agencies includes distributed cognitive systems and actor networks, but also individuals interacting with a single artifact. To make his case, he introduces a number of distinctions. Methodological individualism claims that actions are done only by humans, whereas extended agency theory claims that actions are done by systems comprising humans and artifacts. Moral individualism claims that only humans are responsible for their actions, whereas joint responsibility theory claims that extended agencies are responsible. He writes: "If moral responsibility for an act lies with the subject that undertakes it, and if the subject incudes nonhuman as well as human beings, then so may moral responsibility" (2009, p. 93). So an extended agency, rather than an individual human agent, is responsible for its actions.

One might wonder whether an extended agency deserves blame or praise? According to Hanson, it does. When an extended agency has done something wrong, we often punish it by breaking up the extended agency. For example, a reckless driver is sometimes suspended of his or her driving license and a child misbehaving with a toy is sometimes punished by taking the toy away. These punishments aim at temporarily dissolving the extended agency. "Thus it is reasonable to understand the punishment as directed against an offending relationship rather than a particular part of it" (2009, p. 95). I think this is a valuable insight, but still only the human part of the extended agency experiences negative consequences of dissolving the relationship: the car or toy do not care whether or not they are being used. So, while I do think extended agencies exist, I do not think they can be punished or rewarded for their actions, only sentient beings can. In such extended agencies, I think, only the central controller (i.e., an embodied human organism) can be held responsible for the actions performed by the extended agency.

To help see why attributing responsibility to extended systems consisting of humans and technology is counterintuitive, consider the following case. In 2012, Apple introduced Apple Maps, a new navigation program replacing Google Maps on Apple devices. Initially, there were some bugs in the system, ranging from improper labelling of places to unmapped roads. For example, drivers heading to Fairbank International Airport in Alaska were instructed to drive onto an airport taxiway located directly across from the runway. Likewise, in Australia, drivers heading to the town of Mildura were given incorrect directions off the highway into Murray Sunset National Park, potentially resulting in dangerous situations. Navigation systems are paradigm cases of cognitive artifacts and as the above examples show, their users often put a high degree of trust in these systems. Nothing bad happened, but it is not difficult to image that something bad could have happened.

Who is responsible in these cases? Apple took responsibility by removing the Senior Vice President of iOS software from his position. This makes sense, given a product he was ultimately responsible for malfunctioned. It also seems reasonable to hold the users accountable for their actions to some degree. One might expect users to have at least a minimal amount of scepticism when their navigation system tells them to drive onto a taxiway at an airport or straight into the desert. The designers and users can be held responsible, but not the artifact. So, attributing responsibility to systems comprising of humans and artifacts seems unintelligible, because artifacts are not intentional agents and cannot experience the consequences of repercussions aimed towards them. They are furthermore unresponsive to threats of punishments.

Likewise, in Hutchins (1995) paradigm example of a distributed cognitive system, namely a team of navigators on a navy ship using artifacts to navigate into harbour, morally undesirable events can happen. For instance, when navigating the ship into harbour, one of the sailors could make a navigational error by not having calibrated his or her alidade, which is a device that allows one to view an object or structure and use the line of sight to perform a navigational task. Due to this error, the sailor provides the navigator with an incorrect distance to the shore, the ship hits a rock and sinks. Navy ships have a military organisational structure. Depending on the particular hierarchal structure, the responsibility may be shared between the sailor, navigator and captain in a way consistent with navy policy. But if a similar accident is caused due to a malfunctioning alidade, then it seems reasonable to neither hold the sailor, navigator, nor captain responsible, as the accident happened beyond human control. Note that in neither of these two situations, are the artifacts in the distributed cognitive system held responsible (See Schulzke 2013 for a similar conclusion in a military context). However, this is not to say that responsibility cannot be attributed to distributed systems. Social groups, for example, can sometimes be held responsible for actions of its constituent members (Coeckelbergh and Wackers 2007; Floridi 2013). Consider again the above mentioned example of a team of engineers discussing whether or not to release a potentially dangerous air-control system. If the system is released and causes harm, then the entire group can share responsibility. But again it is plausible to only hold the human agents, not the air-control system, responsible.

Conclusion

This paper compared distributed cognition and distributed morality theory and pointed out a number of areas where these two views overlap and differ in their ontological commitments. Specifically, I pointed out a number of similarities and differences regarding the ontological status of artifacts and the larger systems they constitute in debates on distributed cognition and distributed morality. I argued that some artifacts, depending on the way they are used, have cognitive and moral status, but lack cognitive and moral agency. I also argued that extended cognitive systems have agency when the artifact is fully transparent and densely integrated into the cognitive processes of its user, whereas distributed systems without central control lack agency. Having evaluated and compared these views, I then continued by focussing on moral aspects of cognitive artifacts. I specifically conceptualised how such artifacts (a) embed and extend moral reasoning and decision-making processes, (b) have a certain moral status which is contingent on their cognitive status, and (c) whether responsibility can be attributed to distributed systems.

References

Brey, P. (2014). From moral agents to moral factors: The structural ethics approach. In P. Kroes & P. P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 125–142). Dordrecht: Springer.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford: Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Coeckelbergh, M., & Wackers, G. (2007). Imagination, distributed responsibility and vulnerable technological systems: The case of Snorre A. *Science and Engineering Ethics, 13*(2), 235–248.

Davies, M. (1991). Thinking like an engineer: The place of a code of ethics in the practice of a profession. *Philosophy & Public Affairs, 20*(2), 150–167.

DeMarco, J., & Ford, P. (2014). Neuroethics and the ethical parity principle. *Neuroethics, 7*(3), 317–325.

Floridi, L. (2013). Distributed morality in an information society. *Science and Engineering Ethics, 19*(3), 727–743.

Floridi, L., & Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379.

Giere, R. (2004). The problem of agency in scientific distributed cognitive systems. *Journal of Cognition and Culture, 4*(3), 759–774.

Giere, R. (2006). The role of agency in distributed cognitive systems. *Philosophy of Science, 73*(5), 710–719.

Hanson, F. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and Information Technology, 11*(1), 91–99.

Harris, C., Keil, P., Sutton, J., & Barnier, A. (2010). Collaborative remembering: When can remembering with others be beneficial? In W. Christensen, E. Schier & J. Sutton (Eds.), *Proceedings of the 9th conference of the Australasian Society for cognitive science* (pp. 131–134).

Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences, 14*(3), 577–598.

Heersmink, R. (2016). Extended mind and cognitive enhancement: Moral aspects of cognitive artifacts. *Phenomenology and the Cognitive Sciences*. doi:10.1007/s11097-015-9448-5.

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.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer–Human Interaction, 7*(2), 174–196.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Hutchins, E. (2014). The cultural ecosystem of human cognition. *Philosophical Psychology, 27*(1), 34–49.

Johnson, D. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology, 8*(4), 195–204.

Johnson, J., et al. (2013). The effects of note-taking and review on sensemaking and ethical decision making. *Ethics and Behaviour, 23*(4), 299–323.

Kaiser, M., Millar, K., Thorstensen, E., & Tomkins, S. (2007). Developing the ethical matrix as a decisions support framework: GM fish as a case study. *Journal of Agricultural and Environmental Ethics, 20*(1), 65–80.

Kirsh, D. (2006). Distributed cognition: A methodological note. *Pragmatics & Cognition, 14*(2), 249–262.

Knappett, C., & Malafouris, L. (2008). Material and nonhuman agency: An introduction. In C. Knappett & L. Malafouris (Eds.), *Material agency: Towards a non-anthropocentric approach* (pp. ix–xix). New York: Springer.

Levy, N. (2007a). Rethinking neuroethics in the light of the extended mind thesis. *American Journal of Bioethics, 7*(9), 3–11.

Levy, N. (2007b). *Neuroethics: Challenges for the 21st century*. Cambridge: Cambridge University Press.

Ludwig, D. (2015). Extended cognition and the explosion of knowledge. *Philosophical Psychology, 28*(3), 355–368.

Magnani, L., & Bardone, E. (2008). Distributed morality: Externalizing ethical knowledge in technological artifacts. *Foundations of Science, 13*(1), 99–108.

Malafouris, L. (2008). At the potter's wheel: An argument for material agency. In C. Knappett & L. Malafouris (Eds.), *Material agency: Towards a non-anthropocentric approach* (pp. 19–36). New York: Springer.

Martin, A., Bagdasarov, Z., & Connelly, S. (2015). The capacity for ethical decisions: The relationship between working memory and ethical decision making. *Science and Engineering Ethics, 21*(2), 271–292.

Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded*. London: Palgrave McMillan.

Menary, R. (2010). Dimensions of mind. *Phenomenology and the Cognitive Sciences, 9*(4), 561–578.

Mepham, B. (2000). A framework for the ethical analysis of novel foods: The ethical matrix. *Journal of Agricultural and Environmental Ethics, 12*(2), 165–176.

Michaelian, K. (2012). Is external memory memory? Biological memory and extended mind. *Consciousness and Cognition, 21*(3), 1154–1165.

Michaelian, K., & Sutton, J. (2013). Distributed cognition and memory research: History and current directions. *Review of Philosophy and Psychology, 4*(1), 1–24.

Peterson, M., & Spahn, A. (2011). Can technological artefacts be moral agents? *Science and Engineering Ethics, 17*(3), 411–424.

Rowlands, M. (1999). *The body in mind: Understanding cognitive processes*. Cambridge: Cambridge University Press.

Schulzke, M. (2013). Autonomous weapons and distributed responsibility. *Philosophy and Technology, 26*(2), 203–219.

Smart, P., Heersmink, R., & Clowes, R. (2016). The cognitive ecology of the internet. In S. Cowley & F. Vallée-Tourangeau (Eds.), *Cognition beyond the brain: Computation, interactivity and human artifice* (2nd ed.). Dordrecht: Springer.

Søraker, J. (2007). The moral status of information and information technology: A relational theory of moral status. In S. Hongladarom & C. Ess (Eds.), *Information technology ethics: Cultural perspectives* (pp. 1–19). Hershey: Idea Group Publishing.

Sterelny, K. (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences, 9*(4), 465–481.

Sutton, J. (2006). Distributed cognition: Domains and dimensions. *Pragmatics & Cognition, 14*(2), 235–247.

Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In R. Menary (Ed.), *the extended mind* (pp. 189–225). Cambridge, MA: MIT Press.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago: University of Chicago Press.

Verbeek, P. P. (2014). Some misunderstandings about the moral significance of technology. In P. Kroes & P. P. Verbeek (Eds.), *The moral status of technical artefacts* (pp. 75–88). Dordrecht: Springer.

Wheeler, M. (2010). In defense of extended functionalism. In R. Menary (Ed.), *The extended mind* (pp. 245–270). Cambridge, MA: MIT Press.

Wilson, R., & Clark, A. (2009). How to situate cognition: Letting nature take its course. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 55–77). Cambridge: Cambridge University Press.