Cognitive Technology and Human-Machine Interaction: The Contribution of Externalism to the Theoretical Foundations of Machine and Cyborg Ethics

Marcello Ienca

ANNALS of the University of Bucharest Philosophy Series

Vol. LXVI, no. 2, 2017
pp. 91–115.
Abstract

Machine ethics is the branch of ethics concerned with the behavior of artificially intelligent systems. Cyborg ethics is the related field of investigation concerned with the ethics of human-machine hybrid systems. While these areas of ethical investigation are experiencing rapid growth urged by disruptive advances in artificial intelligence, robotics and human-machine interaction, yet their theoretical foundations continue to elude consensus among researchers. In fact, most attention in machine and cyborg ethics has been devoted to normative and applied ethical questions concerning the moral status of artificially intelligent systems, the moral permissibility of their application in specific contexts, and the normative principles governing the interaction between artificially intelligent systems and humans. While cyborg ethicists have discussed the ethical implications of integrating man and machines, machine ethicists have long debated on whether artificially intelligent systems have the cognitive capacities necessary for the attribution of moral status. It remains unexplored, however, what theory of cognition is best placed to explain and assess these cognitive capacities or competent actions, especially in relation to human-machine interaction. This contribution aims at harmonizing the theoretical foundations of, respectively, machine and cyborg ethics and argues that an externalist account of cognition based on the notion of extended mind might offer a valid substrate for such harmonization.

Keywords: machine ethics, theoretical foundations, extended cognition, embodied cognition, externalism, artificial intelligence.

1. Cognition and the Problem of Moral Status

Machine ethics is the branch of ethics concerned with the behavior of artificially intelligent systems. A basic problem in machine ethics is
determining what types of artificially intelligent systems possess the cognitive capacities necessary for attributing moral status and moral responsibility. The answer to this problem is strictly dependent on another problem, often addressed in cognitive psychology and theoretical artificial intelligence, namely that of determining which mechanism realizes cognition in living organisms and artificial systems. In fact, in order to determine which systems possess capacity X (where X = cognitive capacity necessary and sufficient for the attribution of moral status and moral responsibility), we first need to explain what X is. In accordance with the requirements of modern science, this explanation should ideally be in functionalistic and mechanistic form.

The default position in neuroscience is that cognitive processes in living organisms are largely implemented by the brain. The reason for that stems from the fact that the brain functions, in all vertebrate and most invertebrate animals, as the center of the nervous system, where center has a twofold meaning. First, the brain is the functional center of information processing, which continuously receives sensory information as input, and then, after rapidly analyzing this information, responds by producing outputs which serve to control virtually all bodily actions and functions. Second, the brain is the necessary component of the nervous system, in the absence of which no computation-like information processing connecting sensory inputs to motor outputs would take place. Given these two characteristics, the analogue of the brain in artificial intelligent systems is often described as a central processing unit (CPU) in a serial processing digital computer.

This idea is well rooted in the observation that the brain of humans is a particularly complex organ. Although it has the same general structure as the brains of other mammals (Azevedo et al. 2009), it is over three times as large as the brain of a typical mammal with an equivalent body size. It has been estimated to contain 50-100 billion (10^{11}) neurons, of which about 10 billion (10^{10}) are cortical pyramidal cells (Herculano-Houzel 2009). Most of the expansion of the human brain with respect to the brain of other mammals comes from the cerebral cortex, a convoluted layer of neural tissue that covers the surface of the forebrain, which plays a key role in putatively cognitive processes such as thinking, reasoning, memory, attention, awareness, language and perception.
These cortex-enabled capacities are usually considered by ethicists as co-determinants of moral status. Therefore, the possession of such cognitive capacities among artificially intelligent systems would justify the attribution of moral status to these artifacts. For example, according to Kant, only beings with the capacity for practical rationality have moral standing (Kant 2002 [1788]). In a similar fashion, Bentham (1823) proposed sentience as a discriminant of morality. Other intellectual capacities that have been proposed as grounding full moral status include intentionality (Quinn 1984, Sullins 2006), self-awareness (McMahan 2002), and future-oriented planning (Singer 1993).

For a theoretically well-founded machine and cyborg ethics, however, it is not sufficient to know what cognitive capacities are associated with moral status. Another foundational question is determining how these cognitive capacities can be realized in humans, machines and interactive human-machine systems. Consequently, developing a consistent theory of how cognition can be realized should cast light on how morality-enabling cognitive capacities can or should be realized in, respectively, humans, machines and interactive human-machine systems. Additionally, addressing the question about the realization of cognitive capacities would provide a more solid metaethical foundation to machine and cyborg ethics. Developing a theory of how morality-enabling cognitive capacities are realized at the functional level would cast light on what cognition means in relation to functions or processes that enable (the attribution of) moral properties.

1.1. Internalism vs. Externalism

According to a position in cognitive science which obtained significant success in the ‘70s and ‘80s, the relation between the brain and the cognitive processes it implements basically resembles the relation between computer hardware and system software: the brain, on the one hand, is hardware, i.e. the physical part of the computer; the mind, on the other hand, is software, i.e. a set of programs and related
data installed in the hardware in order to provide instructions for the hardware to accomplish tasks (Block 1995).3

The traditional view regarding the role of external (e.g. bodily or ecological) factors in cognitive processes admits that these factors play a causal role in determining which input patterns will be processed by the nervous system, in particular by the brain, through a finite number of internally defined successive states and manipulated to produce an output. However, this view does not attribute to environmental factors any constitutive role in the information processing itself. According to the traditional view, indeed, cognitive processing in biological organisms corresponds to information processing in the form of electrochemical signaling within the neural circuits of the nervous system.

In the past 20 years, findings in various areas of the cognitive science including cognitive psychology (Provins 1997, Pfeifer and Bongard 2007, Gigerenzer and Selten 2011, Scott et al. 2001, Sparrow et al. 2011), biolinguistics (Bickerton 2009, Fitch 2011, Jenningis 2011), artificial intelligence (Steels and Brooks 1995, Chrisley 2003) and philosophy of mind (Clark and Chalmers 1998) have supported the idea that the functioning of cognition might intimately depend on external (e.g. bodily and environmental) resources. In particular, these findings have led some cognitive scientists to formulate two related (even though not self-implicating) hypotheses: (i) that cognitive processes in humans, non-human animals, AI systems and cyborgs (defined as human-machine integrated systems) might be actively influenced by external (e.g. bodily and environmental) factors (an hypothesis known as ‘embedded cognition’), and (ii) that cognitive processes might partly capitalize on (indeed, be partly constituted by) information being stored and elaborated in natural or artificial representational systems located outside the organism or artificial system (an hypothesis known as ‘extended mind’). These findings, as well as their conceptual implications, have questioned the default brain-centered paradigm of neuroscience according to which

---

3 In recent times, the brain-computer metaphor has become increasingly controversial as many authors have criticized this analogy based on anti-representational approaches to cognitive science. See, among others, Epstein (2016).
cognition is computation-like information processing which is exclusively realized inside the brain.

In this contribution, I will describe this externalist account of cognition by presenting its major theoretical components: embodied cognition, ecological cognition, distributed cognition and situated artificial intelligence. Subsequently, I will argue that this family of externalist approaches might offer a viable contribution to the theoretical foundations of machine and cyborg ethics in the era of human-machine interaction.

2. Forms of Externalism

2.1. Embodied Cognition

Embodied cognition is the view according to which “cognition is deeply dependent upon features of the physical body of an agent, that is, when aspects of the agent’s body beyond the brain play a significant causal or physically constitutive role in cognitive processing” (Wilson and Foglia 2011).

For example, findings in the cognitive psychology of perception (Balcetis and Dunning 2007) have shown that internal bodily states affect distance perception. Participants were randomly assigned by the researchers to three groups: high-choice (or freedom of choice), low-choice (experimenter choice), and control conditions. They were also asked to walk across a certain area. At the conclusion of the experiment, each participant was asked to estimate the distance she walked. The results showed that the high-choice participants perceived the distance walked as significantly shorter than participants in the low-choice and control groups, even though they walked the same distance. These results show the ability of internal states to influence perception of physical distance moved. This illustrates the reciprocal relationship of the body and mind in cognitive processing. Similarly, findings in the cognitive psychology of vision (Bekkering and Neggers 2001) have shown that bodily orientation can affect information processing in visual search, thus supporting the view that vision is often action-guiding, and
bodily movement and the feedback it generates are more tightly integrated into, at least, some visual processing than it was anticipated by traditional models of vision (O’Regan and Noë 2001). Experimental findings in support of the causal role of bodily factors in cognitive processing have been reported with regard to many other cognitive processes, including memory (Scott et al. 2001), language (Olmstead et al. 2009, Lakoff and Johnson 1980), and moral cognition (Greene and Haidt 2002, Haidt et al. 1993, Niedenthal et al. 2005). Such findings have also encouraged AI researchers to bring embodiment theory into Artificial Intelligence. The resulting approach is called Nouvelle AI. Whereas traditional AI (sometimes referred to with the acronym GOFAI, Good Old-Fashioned Artificial Intelligence) has by and large attempted to build disembodied intelligences whose only way of exhibiting human-like cognitive performances is to process symbolic information (regardless of the morphology of the robot that these processes are implemented by), Nouvelle AI attempts to build embodied artificial intelligences in which structural and morphological factors of the robot play a causal role in driving cognitive processes. To the representational and symbolic stance of traditional AI, which was ultimately aimed at simulating human general intelligence, Nouvelle AI approaches oppose an embodied stance aimed at emulating the behavior of evolutionarily simpler organisms such as insects.

2.2. Ecological Cognition

Research in embodied cognition has shown that bodily factors external to the nervous system of a living organism can play a significant causal or physically constitutive role in cognitive processing. Therefore this approach has extended the boundaries of cognition from the nervous system of a living organism to its entire body. The ecological cognition approach attempts to further expand the class of factors that are causally relevant for cognition as to include factors localized outside the organism, i.e. in the local environment where the organism lives and with which it interacts. According to this view, cognition is not exclusively realized by the brain but might, under
certain conditions, emerge at the interplay between the brain, the rest of the body and the external environment. This view is succinctly captured by James Gibson’s motto: “Ask not what’s inside your head, but what your head’s inside of” (Mace 1997).

2.3. Distributed Cognition in Human-Machine Interaction

According to the distributed cognition approach, cognitive information-processing does not emerge exclusively within an individual cognizing agent but is distributed across a cognitive continuum involving the agent and the physical or social structures with which it interacts (Hutchins 1995). Accordingly, a cognitive state, say a perceptual state, is a distributed state that includes the perceiving actor as well as elements in the perceiving actor’s physical and social environment (Gibson 1966, Dror and Harnad 2008).

Distributed cognition has turned out to be a useful approach for analyzing social aspects of cognition (socially distributed cognition) as well as cognizing in the digital technology era. Cognizers in a digital environment tend to offload some of their cognitive functions onto cognitive technologies such as personal computers and the internet, thereby extending their performance capacity beyond the limits of their own brain power (Dror and Harnad 2008). A prime example of cognitive technology is search engine technology. Sparrow et al. (2011) have shown that if people rely on internet-stored information which they expect to be accessible at a later time, they are worse at remembering the actual trivia, but better at remembering where to access it. These results suggest that (i) processes of human memory adapt to the digital environment of computing and communication technology; and that (ii) external cognitive technologies do not merely determine instrumental and quantitative changes, but can rather have qualitative effects on how information is processed. In the light of the two considerations,

---

4 While Dror and Harnad’s research on cognitive technology emphasizes the causal role of external artifacts in cognitive processing, these authors surprisingly reject ‘extended’ approaches to the ontology of the mind. See Dror and Harnad (2008, p. 2).
proponents of the distributed cognition approach think that such types of phenomena are better understood by redesigning human cognition as not confined to the individual cognizing organism but distributed across the network composed by the organism and its (digital and social) environment. In this context, distributed means that the operation of the cognitive system involves (i) various internal and external components and (ii) a functional coordination between these components.

Today, a number of cognitive technologies are available for supporting cognitive processing among people with cognitive disorders or disabilities. For examples, an increasing number of intelligent devices is being developed for providing external cognitive assistance (especially memory assistance in the form of adaptive prompts and reminders) to people with Alzheimer’s disease and other dementias (Ienca et al. 2017). Cognitive technologies do not include exclusively digital artifacts but brain-dependent cognitive faculties too. For instance, language itself is often understood as a form of cognitive technology that (i) allows cognizing organisms to offload some of their cognitive functions onto the brain of other cognizers (social environment), and (ii) extends organisms’ individual and joint cognitive performances, distributing the load through interactive cognition (Dror and Harnad 2008).

### 2.4. Situated Artificial Intelligence

The role of external factors in driving cognitive processes has also been highlighted by recent research approaches in Artificial Intelligence (AI). For example, the situated approach in contemporary AI is aimed at designing artificial agents that are situated in a given environment and are capable of behaving successfully in it. The cognitive architecture of situated artificial systems is commonly referred to as a ‘subsumption architecture’ (Brooks 1986). Whereas classical architectures for artificial systems rely on a central processing unit (CPU), i.e. a hardware unit that carries out the instructions of a program by processing the basic operations of the system through serial processing, subsumption architectures have multiple parallel computing elements, with no one unit considered the ‘center’, and process the information by a
distributed interconnected set of processors. Each processor is specified as a layer of networks of augmented finite state machines (Brooks 1991). Such architecture implies that the cognizing agent does not rely on an internal, symbolic description of the environment, but rather on a non-representational model of the interactions between the agent and its local environment. Simulating artificial agents in a natural or virtual environment requires AI loops, i.e., simulation technologies of the entire process that goes from perceiving an environmental stimulus to an action on the environment. The role of external factors in driving cognitive processes implemented by situated robots is twofold:

1. They causally affect the robot’s sensory system without involving intermediate levels of representation, thus influencing the robot’s internal information processing and behavior.
2. They can be manipulated or modified through the robot’s behavior, for they are linked with the robot in a dynamic interaction loop.

The most significant characteristics of situated artificial systems as compared to classical AI are (i) a refined internal organization, in particular in terms of computational cheapness and information-processing speed; and (ii) a refined capacity of behaving and acting in a dynamic environment. Similar characteristics can be detected also in environment-dependent cognizing in living organisms. For instance, heuristic-based decision-making has turned out to be computationally less expensive and to produce faster and more accurate behavioral patterns than classic computational decision-making (Gigerenzer and Selten 2002).

2.5. At the Origins of Cognitive Externalism: Evolutionary Hypotheses

Characteristics like optimization and output accuracy have led authors such as Rowlands (2003), Clark (2002, 2008), and Gigerenzer (2007) to advance the hypothesis that embodied and ecological cognition are adaptations, i.e., phenotypical traits evolved by natural selection. Geary (2005) and Striedter (2005) have set a list of adaptive criteria that a certain cognitive faculty should meet in order to make its selection evolutionary predictable. These include, among others, optimization in
system internal organization, optimization in input processes, and positive feedback on other system faculties.

Internal organization is central not only in the assessment of the evolutionary predictability of a cognitive process but also in the assessment of its functional organization. From an evolutionary perspective, both in the sense of evolutionary biology and evolutionary computation, the ability to extend some cognitive processes to the external environment might determine an optimization in our internal system organization in three important ways. First, it may produce a better metabolic equilibrium. Cognitive systems are dissipative systems that get pushed into operation by harnessing energy from a variety of metabolic pathways. The human brain, in particular, is very dissipative for it claims only 2% of the body mass, but is responsible for approximately 20% of our body oxygen consumption (Rolfe and Brown 1997). For a cognitive system, therefore, energy must be constantly available for work (e.g. mechanical work) or for other processes (such as chemical synthesis and anabolic processes). However, energy is not always easily available for a system. Food, for instance, humans' best resource to assimilate some of the essential nutrients that our cells convert in energy, is often scant. For this reason, authors have argued that evolution might have favored those organisms capable to spark their life-maintaining processes with the lowest possible expenditure of chemical energy. According to a principle in bioenergetics, all living systems try to execute their biological processes with the smallest effort/profit ratio, namely to obtain the best possible outcome with the lowest possible energy expense (Morovitz 1999). One possible way for the nervous system to reduce such effort/profit ratio might have been by transferring some processes from neurons or single processing units to external resources, as the latter do not draw on internal energy supplies.

Optimization of the internal organization does not operate exclusively at the biochemical level, but at the functional level too. In order for a system to be functionally optimized in evolutionary terms, and thus to have a high statistical probability of propagating itself to future generations, it must be able to (i) execute more functions than its unoptimized matching system; (ii) execute the functions of its unoptimized matching system more efficiently. In the case of
environment-dependent cognition both conditions seem to be satisfied. In the first place, through extending to the external environment, the cognitive system might be able to execute more cognitive functions than if it were confined within the original boundaries. This functional advantage might not only pertain to couplings with sophisticated technologies, but also to simple artifacts and even to parts of the physical body too. For instance, McClelland (1989) and Clark (1989) observed that, thanks to the use of pen and paper, students can perform complex arithmetical and geometrical operations that they could not solve if they would only lean on internal resources. The same goes for children counting fingers on their own hand (see Ginsburg 1989, Dehaene 1999, Carpenter et al. 1999) as well as for many non-human species. In addition, extended systems have sometimes been observed to be more efficient than non-extended ones, as they are able to process information faster and to produce more accurate outcomes. Kirsh and Maglio (1994), for example, calculated that the physical rotation of a shape in the computer game Tetris goes about three times faster than the mental rotation of the same shape, precisely 300 milliseconds of 1000 milliseconds to rotate the same shape through 90°. This reveals external processing to be, at least under some conditions, dramatically faster than internal processing. The same can be said, again, for mathematical operations. If one compares the performance of mathematical exercises both with and without a calculator (or pen and paper set), one would suddenly notice a dramatic difference in the time it takes to work them out.

In addition, leaning on external supports might not only increase the processing speed, but the outcome accuracy too. This phenomenon is particularly common among people with cognitive disorders, especially older adults with dementia. In fact, deficits in the accuracy of beliefs caused by insufficient internal cognitive resources have been observed in patients with Alzheimer’s disease and frontotemporal dementia (Fernandez-Duque et al. 2008). In these patients, memory reduction caused by loss of neurons and synapses in the cerebral cortex systematically leads to a deficit in belief accuracy. For example, a patient more prone to forget, say, the name of his daughter Amy, will also be more prone to have the false belief that his daughter’s name is Laura.
2.6. The Extended Mind

Extended Mind (hereafter EM) is the thesis, first proposed by Clark and Chalmers (1998), according to which the mind should not be limited to internal information-processing in the nervous system of cognizing organisms (or internal hardware of an artificial system) but extended to include some functionally isomorphic processes whose local position is outside the nervous system, and even the body of cognizing organisms, or internal hardware of an artificial system (Rowlands 2003). More succinctly, EM is the view according to which the mind of a biological organism or artificial system may be (partly) constituted by extended cognitive components located outside that organism or system.

From the perspective of EM, the only parameter that defines the components of a cognitive process is the act of playing a constitutive role in a cognitive network, regardless of whether these components are physically located inside or outside the organism or artificial system (Chalmers 2008).

Not all external (bodily, technological and environment-dependent) processes count as mind-constituent. Rather, in order to be considered extensions of the mind, external processes must satisfy two basic conditions: (i) functional equivalence, and (ii) reliability of coupling. Let us see what this means.

According to a basic principle of EM, called the ‘Parity Principle’ (Clark and Chalmers 1998), if an external system performs a process functionally equivalent to a process that (i) could be executed by an internal cognitive system, and (ii) if executed by an internal cognitive system would be regarded as cognitive without doubt, than the external process should be regarded as cognitive as well. Therefore, according to this view, cognition is not bounded by the nervous system of the organism, but may extend into processes that are realized by systems physically located outside that organism. The boundary of the nervous system becomes thus arbitrary and explanatory vacuous in determining the boundary of the implementation medium of cognition, as cognition might be implemented by entities located outside the nervous system of the organism. As Clark and Chalmers (1998) famously put it:
If, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing as part of the cognitive process, than that part of the world is (so we claim) part of the cognitive process. Cognitive processes ain’t (all) in the head! (3).

The reliability of coupling criterion introduces a further restriction to the EM cases. According to this criterion, not all external processes that are functionally equivalent to internal cognitive processes should be regarded as mind-constituent. Rather, functionally isomorphic processes are considered mind-constituent only if they couple with the internal cognitive processes that they are isomorphic to in a reliable way (Clark and Chalmers 1998). Reliability is a complex property. In the EM debate it is commonly thought to involve the following sub-properties: availability, portability and design.

Availability requires the external process to be coupled with the internal one in a manner that the external process can be easily and quickly accessed by the internal system. An example of an external process that meets this requirement is the use of smartphone technology, as these external resources provided by these handheld devices are usually constantly available to the cognizing human being during her every-day-life.

Portability requires the external process to be coupled with the internal one in a manner that is easily transferable in space and over time. An external process is said to be portable if it does not get decoupled when the internal system changes its local position in space nor when minor alterations affect the local environment itself. Sun-based communication in honey bees meets this requirement too, as the sun position continues to be coupled with the sun also when the bee changes its local position. A human-scale example is the use of wearable technology (e.g. smart watches), as these resources can be taken with the cognizing human beings who wear them in a more reliable and robust manner compared to other device types such as desktop computers. In fact, their portability is guaranteed even in case of macroscopic changes in the user’s local position or minor environmental changes (e.g. seasonal change or change in weather). It is worth considering that portability is not an all-or-nothing property but a continuum. For
example, handheld devices are more portable than desktop computers, wearables are more portable than handheld devices and implants are more portable than wearables. The element of portability is particularly important for assistive technologies developed for people with memory impairments. In fact, these patients often forget to bring their devices with them, hence need devices that exhibit a high degree of frictionless portability.

To these two criteria, which are widely discussed in the literature, I add a third requirement, *i.e.* design. Design requires that the internal-external coupling is not random, but designed to execute the cognitive function that it actually executes. For instance, hardware and software interfaces are non-random in the strong sense that they are designed, built and programmed by technicians precisely to execute the function that they actually execute. Similarly, smartphone technologies are designed, built and programmed precisely to provide an accessible, portable and useful support to internal cognition in humans. Of course, in order to be designed, external objects do not necessarily need to be artifacts or manipulated objects, such as tools. Non-manipulated objects in the natural environment can be also regarded as cognitive extensions, as long as they are co-opted by the cognizing organism for the function they actually execute in the integrated cognitive loop. This cooptation is multiply realizable, as it does not necessarily involve the physical modification of the object, but simply a change of function. For instance, the sun is coopted by the honey bee as vehicle of meaning regarding the position of food resources. Similarly, a tree can be coopted by humans as an external memory support (*e.g.* as a path-tracker sign).

Some authors (Clark and Chalmers 1998, Rowlands 2010) suggest a stronger criterion of design. According to this stronger criterion it is not sufficient that the external system is designed by the internal system to be coupled with it and execute the function it actually executes in the integrated cognitive loop. Rather, the internal system should also be designed to be coupled with external systems and integrate them in the cognitive loop. In other words, the dependence relation between the internal and the external system should not be unidirectional (from the internal to the external system) but bidirectional (from the internal to the external system and the other way round). This stronger criterion of design is easily satisfied by artificial agents, such as situated intelligent
robots. These agents do not simply redesign external objects in order to co-opt them for cognitive functions useful to the robot and couple them with the internal system, but they are themselves designed, built and programmed to couple with external objects and co-opt them for cognitive functions useful to the robot (Hendriks-Jansen 1996). According to Rowlands (2010) the strong criterion of design is satisfied by living organisms too. Based on the evolutionary explanations summarized in the previous section, these authors claim that the cognitive systems of living organisms are designed by (in the sense of ‘selected for’) natural selection to couple with external objects in the environment and co-opt them for cognitive functions useful to organism. This conjecture is logically linked to the empirical hypothesis of environment-involving processes to be adaptations. If this hypothesis turns out to be empirically correct, then cognizing living organisms satisfy the strong criterion of design too.

As we have seen, the reliability of coupling is a crucial criterion to provide EM with a valuable ontology of the mind. In the absence of it, mind-attribution would be ubiquitous (Clark and Chalmers 1998, Clark 2002).

3. Extended Mind as a Theory of Human-Machine Interaction

EM is particularly appealing to describe the class of cognitive interactions between humans and technological devices, hence the behavior of human-machine integrated systems. One classic example that EM proponents usually refer to is the every-day use of smartphones and personal computers to store and retrieve or simply access information. For instance, consider the case of a market trader saving on her smart-phone’s phone-book her customers’ phone numbers, a child searching on a web dictionary the meaning of the word “idiosyncrasy”, a student using a calculator to do her math homework, a teen googling the lyrics of a song she can’t recall, a teacher using an electronic calendar to memorize the course program, a tourist using geolocation to find her hotel in Paris or a virtual translator to communicate with her French waiter. In all these cases, the internal cognitive system of the subject establishes an interaction with some external electronic devices that play
an active causal role in driving a certain cognitive process—respectively, in the cases mentioned above: memory storage, recall, learning, calculation, spatial navigation, language. Moreover, in all these cases the process executed by the external device is functionally equivalent to a process, or at least a phase of a process, executed, or that could be executed, by an internal cognitive system such as the human nervous system. Finally, the coupling between the internal and the external system satisfies the condition of reliability, which is a critical requirement for the successful integration of human and machine. Other classical examples focus on the role of non-electronic technologies and cultural artifacts such as writing instruments, books, codes, etc.; and bodily components such as the hands when they are used to support mental calculation. From the point of view of EM, when people count along their fingers (process known as dactylonomy) there is no theoretical impediment to claim that this external calculation becomes part of a broader calculation process executed by the extended cognitive system composed by the nervous system and the hand. Similarly, when people use notebooks, books and encyclopedias to store or recall information, these external artifacts are regarded as processing units of an extended memory process (Sparrow et al. 2011).

One further privileged field of application of EM are clinical cases of impaired general or modular cognitive ability (Clark and Chalmers 1998). Dementia patients, such as Alzheimer’s disease or vascular dementia patients, tend to compensate for their loss of cognitive ability by using external resources. For example, Alzheimer’s patients in the early and moderate stage of the disease, typically display cognitive inabilities such as the inability to build new memories or to recall vocabulary (Förstl 1998). For this reason, they tend to note down the information that their brain is unable to store and recall it via external resources such as electronic devices or simple notebooks. Today, a broad spectrum of intelligent assistive technologies is available to provide and enhance such external cognitive assistance (Ienca et al. 2017). From the point of view of EM, these external tools get integrated by the cognizing subject in such a way that they literally become constitutive components of the cognitive system responsible for the execution of the cognitive
task. In the case of tool-using Alzheimer’s patients, therefore, cognitive processes such as forming new memories or recalling vocabulary are not exclusively realized by the nervous system, but extend into the external resources exploited by the patient. In fact, the way intelligent assistive devices and notebooks perform (stages of) the memory process in Alzheimer’s patients is functionally equivalent to cognitive processes that, in normal subjects, are usually executed by the brain (Clark and Chalmers 1998).

4. Externalism as a Framework for Cognitive Science and Artificial Intelligence

In virtue of its high level of generalization across the human-machine continuum, the various externalist approaches described in above, have been proposed as a new paradigm or theoretical framework for the cognitive sciences (Clark and Chalmers 1998, Rowlands 2010). The reason for that stems from the fact that these approaches appear particularly suitable for integrating and making sense of the body of evidence in cognitive science and artificial intelligence regarding the role of external factors in cognition. In addition, they offer a suitable common ground for the many areas of cognitive science including cognitive neuroscience, cognitive psychology, AI, cognitive linguistics and philosophy of mind. As a broad theoretical framework, externalism encompasses a large constellation of theoretical and empirical perspectives, which all recognize the causal (and, in the case of EM, even constitutive) role of external resources in driving, supporting or enhancing the internal cognitive capacity of biological organisms.

From a theoretical point of view, externalism appears to unify virtually all approaches and perspectives to the study of cognition that do not underestimate the causal role of environmental factors in driving cognitive processes. In particular, externalism is thought to integrate the embodied cognition approach, as, according to EM, cognition (particularly cognition that leads to competent action in the world) is not confined

---

5 For a detailed analysis of the criterion of constitution see Palermos (2014).
within the brain but extends to components whose local position is outside the brain and even the entire nervous system of a cognizing organism. Since the embodied cognition approach assumes cognition not to be restricted to information-processing in the brain but to extend to bodily components such as the musculoskeletal system and the sensory-motor mechanisms that are external to the organism’s nervous system, then embodied cognition is inherently externalist. More specifically, certain accounts of embodied cognition qualify as special instances of EM if such bodily factors are constitutively relevant to cognition.

In addition, externalism is able to integrate all approaches and perspectives that recognize the causal role of external factors in driving cognitive processes. In particular, externalism is able to integrate the embedded, the situated and the distributed approach to cognition. All these approaches assume cognition not to be restricted to information-processing within the cognizing organism but to extend to processes partially implemented by objects in the organism’s local environment, or at least to emerge out of the interaction between the organism, the intelligent machine and their local environment.

5. Externalism as a Framework for Machine and Cyborg Ethics

While there is a growing consensus that various forms of externalism might offer a valuable framework for the cognitive sciences, it is less intuitive to see how externalist approaches might valuably contribute to strengthening the theoretical foundations of machine and cyborg ethics.

As I have previously stated, machine ethics is the branch of ethics concerned with the behavior of artificially intelligent systems. This discipline deals with problems such as determining the moral status of intelligent machines and designing machines that exhibit moral behavior. I argue that an externalist approach to cognition can provide a scientifically informed and philosophically innovative substrate for addressing these ethical questions. The reason for that stems from a twofold consideration.
First, while many authors recognize that cognitive faculties are critical for the attribution of moral status and moral responsibility (Kant 2002 [1788], Bentham 1823, Quinn 1984, McMahan 2002, Singer 1993, Sullins 2006), yet the field of machine ethics lacks both (i) a uniform meta-ethical theory of what cognition is and (ii) a theoretical characterization of how cognition is realized in intelligent machines. I argue that the externalist framework is epistemologically well-equipped to fill these gaps in the foundations of machine ethics.

From an externalist perspective, cognition can be best understood as information processing that emerges out of the interplay between a cognizing agent (human or intelligent machine) and its physical, digital or social environment. This meta-ethical characterization is general enough to encompass both human and machine cognition, hence can harmonize the meta-ethical foundations of machines ethics with those of moral psychology and applied ethics. This also emphasizes how moral status is supervenient on the level of complexity of cognitive systems, i.e. it is a variable that can be gradually increased or decreased on a continuous scale based on the cognitive system’s degree of complexity.

Additionally, since externalism admits that cognition is realizable through distributed systems consisting of various cognitive components (extended cognition), it provides a theory of realization that accounts not only for standalone intelligent agents but also for parallel and distributed systems. This characterization is particularly fruitful to frame the debate over the moral status and responsibility of emerging trends in computing and AI such as parallel and distributed computer systems (consisting of various networked and communicating components), distributed artificial intelligence and multi-agent systems (i.e. systems composed of multiple interacting intelligent agents within an environment). These types of intelligent machines, in fact, in virtue of their distributed organization, are not accounted for by meta-ethical theories that define cognition only in terms of mental representations implemented by an internal processor (e.g. the human brain or equivalent in silico).

Second, externalist approaches provide a fruitful and informative scientific substrate for designing and developing intelligent machines that exhibit moral behavior. While traditional approaches to the design
of moral machines have primarily focus on producing moral behavior by intervening on the internal cognitive resources, externalism poses the accent on the interplay between the machine and its environment. This theoretical shift and expansion is particularly valuable to account for intelligent machines such as social and companionship robots as well conversational agents. In fact, in these types of machines, the ability to competently interact with the social, digital and physical environment is absolutely critical. This increased attention on the interactive and interpersonal dimension is particularly valuable for social and assistive robots used for assisting frail seniors or people with cognitive disabilities as these patients are often vulnerable individuals, hence ensuring the moral behavior of care robots in an interactive dynamics is a priority (Ienca et al. 2016).

With regard to cyborg ethics, the epistemological advantage of shifting the focus on extended cognitive networks is even greater. The reason for that stems from the fact that cyborgs are, by definition, cognitive agents consisting of both organic and mechatronic components. Therefore, the field of cyborg ethics is highly in need of a theoretical foundation that can account for the entire bio-mechatronic continuum, without arbitrary restrictions in the attribution of moral status and responsibility based on the type of physical realization of cognition within each component. Externalist approaches to the theoretical foundations of cyborg ethics have the advantage of providing a solid and comprehensive foundation to the ethics of all human-machine integrated systems, regardless of the physical realization of their components. Instead of focusing on the hardware architecture of each part of the cyborg, they can provide a comprehensive framework that encompasses the entire human-machine continuum. This epistemological shift is particularly important to account for the increasing use of integrated assistive technologies among people with physical or psychological disabilities such as brain-computer interfaces and neural prosthetics. Additionally, it can account for emerging phenomena where the relevant cognitive processing occurs across a reliable coupling of human and machine (e.g. information search across the brain-smartphone continuum).
Additionally, at a more practical level, authors have noted that an externalist approach to the theoretical foundations of cyborg ethics can broaden our normative conception of harms to technological equipment and provide increased legal protection in the era of human-machine interaction (Carter and Palermos 2014). In their view, to the extent that externalist approaches are prioritized, intentional harm towards technological devices that have been appropriately integrated, should not be simply regarded as property damage, but as "extended personal assault" (ivi). In fact, the role played by these technologies within the extended cognitive process initiated by the user is such that their damage by malevolent third parties might qualify as personal assault. This normative advantage is particularly helpful to account for emerging dual-use risks in cognitive technology, such as malicious brain-hacking (Ienca and Haselager, 2016). With the pervasive diffusion of intelligent computing and its progressive integration into human life, ethics is increasingly required to provide a coherent normative ground to orient society across this historical transformation. Externalism is well positioned to accomplish this task and harmonize not only the theoretical foundations of, respectively, machine and cyborg ethics, but also the interoperability of these theoretical foundations as part of normative human-machine continuum. Authors have observed that technological innovation at the human-machine interface urges a new ethics of the post-human or "more-than-human" moral world (Torrance, 2011). As the boundaries between humans and machines progressively blur, theories of cognition and moral status that rely on internalist, realization-specific accounts slowly become explanatory inadequate. In contrast, externalist approaches enable a shared and common grounding that accounts for all components and modes of realization of the human-machine entanglement.

6. Conclusions

Although machine and cyborg ethics are experiencing rapid growth – urged by disruptive advances in artificial intelligence, robotics and human-machine interaction, yet their theoretical foundations remain undefined or
even affected by conceptual muddles. In fact, most attention in machine and cyborg ethics has been devoted to normative and applied ethical questions concerning the moral status of artificially intelligent systems (Anderson 2013), the moral permissibility of their application in specific contexts (Matthias 2015), and the normative principles governing the interaction between artificially intelligent systems and humans (Sharkey and Sharkey 2012). Machine ethicists have largely discussed whether artificially intelligent systems have the cognitive capacities necessary for the attribution of moral status as well as whether these systems are able to perform competent actions. However, it remains unclear what theory of cognition should better explain and assess these cognitive capacities or competent actions, especially in relation to human-machine interaction. This contribution has described an account of cognition in artificially intelligent systems and living beings from an externalist perspective.

Given its capacity to explain cognition across the entire human-machine continuum, this externalist account of cognition provides a viable foundation for machine and cyborg ethics. In providing an elegant theory of how cognition is implemented in biological and artificial systems, this account also provides a more solid meta-ethical description of what cognition means in relation to functions or processes that enable (the attribution of) moral properties in artificial or hybrid intelligent systems. Functions or processes that enable (the attribution of) moral properties, in fact, can be realized not only via biological organisms (such as humans) but also by artificially intelligent systems and human-machine integrated systems. For this reason, the externalist account of cognition proposed by advocates of the extended mind thesis offers a more suitable and epistemically informative foundational framework for machine and cyborg ethics.

REFERENCES


