ARTICLES

Intelligence, Artificial and Otherwise

Paul Dumouchel

Abstract The idea of artificial intelligence implies the existence of a form of intelligence that is "natural," or at least not artificial. The problem is that intelligence, whether "natural" or "artificial," is not well defined: it is hard to say what, exactly, is or constitutes intelligence. This difficulty makes it impossible to measure human intelligence against artificial intelligence on a unique scale. It does not, however, prevent us from comparing them; rather, it changes the sense and meaning of such comparisons. Comparing artificial intelligence with human intelligence could allow us to understand both forms better. This paper thus aims to compare and distinguish these two forms of intelligence, focusing on three issues: forms of embodiment, autonomy and judgment. Doing so, I argue, should enable us to have a better view of the promises and limitations of present-day artificial intelligence, along with its benefits and dangers and the place we should make for it in our culture and society.

Keywords AI; analytical agents; autonomy; embodiment; individuals; intelligence; judgment; responsibility

✍ Paul Dumouchel, Ritsumeikan University, Graduate School of Core Ethics and Frontier Sciences, 56-1 Kita-ku, Kitamachi, Kita-ku, Kyoto 603 8577 Japan ✉ dumouchp@ce.ritsumei.ac.jp iD 0000-0002-6979-3665

Forum Philosophicum 24 (2019) no. 2, 241–58 Subm. 6 December 2019 Acc. 9 December 2019
ISSN 1426-1898 e-ISSN 2353-7043 DOI:10.35765/forphil.2019.2402.11

What is (artificial) intelligence?

Human intelligence is notoriously hard to define. What does it mean to say that an answer is intelligent, or that a person is brilliant, apart from the fact that the answer is unexpected and striking, or that the person who found it had access to only a small part of the relevant information? Intelligence seems to fall within the category of those things of which we say "I can recognize it when I see it, but would not be able to define it." Perhaps as an attempt to overcome this difficulty, psychologists have devised many tests that aim at measuring intelligence. However, it is not entirely clear exactly what these measure. They have often been accused of bias, of favoring or discriminating against different racial or cultural groups, against women, or against those who are handicapped. Over many years, the highly polemical debates surrounding this have resulted in a proliferation of such tests, with different ones aiming to measure different forms or aspects of intelligence. In every case, what these tests target are various abilities, including the ability to reason, to understand, to store information, to analyze, to synthesize, to retrieve information, and to process auditory or visual stimuli. They measure the speed of processing and decision time, reading and writing ability, quantitative reasoning, and short and long-term memory. Intelligence so conceived would appear to be reducible to a collection of loosely related cognitive abilities, yet the tests are such that together they yield a unique measure, the IQ level, which is viewed as being related to certain types of behavior or conducive to particular valued social ends. For example, variance in IQ is correlated with, among other things, income, job performance, academic success, criminality, juvenile delinquency, teenage pregnancy, and so on.

"Intelligence" as applied to humans, then, is certainly not a well-defined term. It does not correspond to a single unified capacity or faculty. Rather, it refers to a cluster of different abilities that are deemed necessary to achieve some desired objective like academic or economic success. However, because we, as individual human beings, are the bearers of this intelligence, and because intelligence tests yield a single measure, it is nonetheless commonly thought of as just one single coherent ability. Furthermore, because the tests provide a unique scale on which to compare different individuals as more or less intelligent, they encourage us to think that intelligence is itself something—a property that exists in itself, independently of the individuals who manifest intelligent behavior.

A somewhat similar, but nevertheless different situation arises in the case of AI. One generally recognized difficulty in this field is the presence of disagreement over what could count as a precise definition of artificial intelligence. A common solution to this is to adopt a broad-ranging, highly inclusive definition. For example, the "One Hundred Year Study of Artificial Intelligence" project at Stanford University characterizes AI as

> a set of computational technologies that are inspired by—but typically operate quite differently from—the way people use their nervous system to sense, learn, reason and take action.[1]

This lack of precision is actually in keeping with the original project of artificial intelligence. When, in 1955, John McCarthy, Marvin Minsky, Nathan Rochester and Claude Shannon introduced the term and proposed it as a new domain of inquiry, their goal was to explore what intelligence is by reproducing with the help of computers its various aspects, such as reasoning, perception, calculation, or memory. They took as a methodological starting point the idea that every aspect of learning, or of any other characteristic of intelligence, can be described so precisely that it becomes possible to build a machine able simulate it.[2] It is therefore hardly surprising that what today constitutes artificial intelligence is, in fact, a disparate and loosely connected set of computation-based technologies that try to "imitate" or "reproduce," or "are modeled on," various abilities that are typically associated with human intelligence. Yet, as in the case of human intelligence itself, people tend to construe artificial intelligence as something that it is simple and unified. Even so, instead of a particular faculty or ability, and perhaps because the "bearers" of that "capacity" are not in this case clearly recognizable individuals, they often view artificial intelligence as a special entity—something that exists in itself: "AI." Since they also consider that intelligence can be measured on a single scale, allowing us to compare and rate as more or less "intelligent" different "bearers" of intelligence, whether they be human or artificial, they conclude that it makes sense to compare and measure the relative "dimension," "force" or "power" of human intelligence and artificial intelligence.[3]

Artificial intelligence, then, like human intelligence, is not a well-defined category. It does not correspond to a single faculty, but to a collection of computational technologies inspired by some human cognitive abilities. It

[1] See, https://ai100.stanford.edu/

[2] Jean-Gabriel Ganascia, *Le mythe de la singularité* (Paris: Seuil, 2017), 75.

[3] Examples of comparative measures of human and artificial intelligence, generally to our disadvantage, abound. See, for example, James Barrat, *Our Final Invention: Artificial Intelligence and the End of the Human Era* (New York, NY: Dunne/St. Martin, 2015) or Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: OUP, 2014).

comprises loosely related tools and methods: machine learning, deep learning, cognitive computing, data science, big data. These technologies not only aim to imitate or out-perform certain natural cognitive abilities, but also have as their goal allowing us to explore various domains of inquiry that would be closed to us in the absence of such tools—either for reasons of speed, or because we face other types of limitations. For example, we cannot directly explore the inside of a brain without destroying it, but artificial cognitive systems, such as fMRI allied to AI (deep learning), allow us to (re)create images from its internal functioning.[4] As with most tools, these ones are explicitly developed to do that which we could not do without them. Just as an airplane or a crane allows us to fly or to lift extremely heavy objects, thanks to artificial cognitive systems we are now able to accomplish tasks and operations which were unthinkable until recently. If computers were as slow and error-prone as we are when calculating, we would not have invented them—or at least, if we had done so, we would not use and rely on them as we now do.

It follows that in comparing human cognitive abilities and artificial intelligence, no general conclusions can be drawn, and, especially, that *there is no universal scale on which we can compare human and artificial intelligence.* Such comparisons are always local and partial. They bear on specific, often highly particular, abilities.[5] It also follows that questions such as "Can artificial intelligence surpass us?" are not well formulated, and thus are highly misleading, inasmuch it is already evident how they are to be answered: of course these artificial systems can do better than us. That, after all, is precisely why we have developed them: to calculate faster than us, to be able to react with greater precision to changes in the environment than we do, to detect and track small variations that we ourselves are unable to perceive. Asking a question whose answer is already obvious can be misleading, because it suggests that the latter sort of answer cannot be the right one, and that something else, something more momentous, must be going on. Yet if there is anything surprising here at all, it is surely just the extent of our

[4] Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani, "Deep Images Reconstruction from Human Brain Activity," *bioRxiv*, accessed October 1, 2019, https://doi.org/10.1101/240317.

[5] This is often evident in the way these comparisons are formulated or reported. It is claimed, for example, that an artificial system can diagnose cancer with as much accuracy as a trained physician, or that it can recognize the presence of a cat in an image with a higher rate of success than an average human. These achievements, however, correspond to highly specific capacities—a much more useful one in the case of cancer-detection than in that of identifying cats—and as such do not constitute forms of general knowledge.

own ingenuity. Such questions are not well formulated inasmuch as they suggest that we are comparing, with reference to a unique scale of gradational distinctions, the same property—intelligence—as manifested in two different but nevertheless equivalent bearers of it: namely, human beings and artificial systems. However, neither human nor artificial intelligence can be said to amount to a coherent and unified faculty, and nor is there any scale that would allow us to compare them *in toto*. This does not mean that there are no important and urgent questions to be asked concerning the place and role of these technologies in our world, but whether or not society will be taken over by some super-intelligent artificial entity that will enslave or even eliminate us is definitely not one of them. Questions such as this are nothing better than "opium for the people," serving as they do to distract us from the real matters of concern.

Once we abandon the fantasy of being able to measure as a whole the "strength" or "power"[6] of artificial intelligence compared to human intelligence, more interesting forms of comparison become possible. In the remainder of this paper, I wish to focus on three issues or aspects of the way intelligence is realized in humans and in many of the artificial systems we now build. The first is *embodiment*. To what extent is the way in which a cognitive system is embodied relevant to what it can (or cannot) do? What consequences follow from the different ways in which human and artificial systems embody intelligence? The second issue is that of *autonomy*. What does it mean to say that an artificial system is autonomous? Is this different from what we mean when we attribute autonomy to humans? In this section, I will also touch upon the question of moral autonomy. Finally, I will address the question of *judgment*. In the context of philosophy, this is an issue not often raised in connection with artificial intelligence. However, it is implicit in much of the literature that concerns itself with the consequences of the proliferation of artificial cognitive systems in such socially required areas as the legal sphere, job-candidate selection, performance evaluation, the screening of social-welfare applications, and so on.[7] I argue that any deficiency or failure at the level of judgment such

[6] The correct formulation would be "to measure intelligence in both human and artificial intelligence," but it is not clear what talk of the "intelligence of artificial intelligence" could mean. This difficulty illustrates my earlier point: namely, that intelligence is neither a "thing" nor a well-defined category.

[7] See for example: Virginia Eubanks, *Automating Inequality* (New York, NY: St. Martin's Press, 2018); Mireille Hildebrandt, *Smart Technologies and the End(s) of Law* (London: Edward Elgar, 2016); Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality* (New York, NY: Crown Books, 2016).

as may be identified as common to different artificial systems constitutes a philosophical issue, and not merely a reflection of the narrow interests of those responsible for developing the systems in question.

As we shall see, the three issues of embodiment, autonomy and judgment are closely related, this being for deep philosophical reasons that have little to do with the debate about mind and matter. The differences between artificial and human intelligence to which I point do not reveal some ultimately unsurpassable limits pertaining to manmade systems. Rather, they reflect how these systems are made, something that partially depends on what we ourselves want from them. Whether, in the future, we might be able to create artificial cognitive systems whose form of intelligence would be closer to human intelligence is, I believe, an open question. My claims should therefore not be construed as reflecting some understanding of the definite characteristics of artificial systems *per se.* Instead, they seek just to illustrate some features of present-day artificial cognitive systems—ones that are important, and which are, I argue, also consequences stemming from our own choosing of certain forms of embodiment and of autonomy for these artificial systems.

Embodiment

Embodiment is now a major topic, and has come to define a significant new approach within, both cognitive science and philosophy.[8] It first appeared as a reaction to and criticism of the view that all cognition is computation. This computational view is implicit in the original project of artificial intelligence. The idea that we can reproduce all aspects of human intelligence with the help of computers implies, firstly, that there is nothing more than computation to human intelligence, as computing is all that a computer itself can do. Secondly, it also implies that whether the system which carries out these computations is a human brain, or an artificial system whose components are silicon chips rather than neurons, is a matter of little or no consequence for what it can come to know or how it arrives at this. Finally, it implies, at least implicitly, that as cognitive beings humans are—or can be reduced to—their brain as the organ that carries out computation, so that nothing else counts.[9]

[8] See, for example, Anthony Chemero, *Radical Embodied Cognitive Science* (Cambridge, MA: MIT Press, 2009); Rolf Pfeifer and Josh Bongard, *How the Body Shapes the Way We Think* (Cambridge, MA: MIT Press, 2007); James Stewart, Olivier Gapenne, and Ezequiel A. Di Paolo, *Enaction. Towards a New Paradigm for Cognitive Science* (Cambridge, MA: MIT Press, 2010).

[9] For an analysis of this ideology—that humans are essentially their brain—and its history, see Fernando Vidal and Francisco Ortega, *Being Brains. Making the Cerebral Subject* (New York, NY: Fordham University Press, 2017).

Embodiment, construed as a research project, rejects these presuppositions. It argues that there is more to cognition than computation, and that we are more than our brains. The body, the material that gives shape and makes real a cognitive system, is not like a dress or suit that the "mind" could change at will, but part and parcel of the system itself. It participates in what and how the agent knows, not as something external to the mathematically (computationally) defined cognitive system, but as what makes that cognitive agent what it is. Embodiment is thus the idea that the material, and not only the computational, dimension of a cognitive system—be it natural or artificial—is fundamental to the way in which it knows and participates actively in such knowing. This contribution at the material level, it is argued, is proof of the fact that there is more to knowledge than mere computation, and that as knowing subjects there is more to us than our brains.

Many artificial cognitive systems, in the sense of artificial agents that can know their environment and react autonomously to changes in the latter, are embodied in a very different way than we humans are. The point I wish to insist on here is not so much the contrast between the silicon-based hardware of computers and the neuron-based wetware of human brains. Embodiment, as I understand it, primarily concerns something else. The body of an agent is not simply the material stuff of which it is made: embodiment also refers to how the agent acts in the world and communicates (interacts) with others. From a philosophical point of view, what this means is that rather than construing the body as a prison for the soul, hiding the latter from the world, we should rather think of it in the opposite terms, as corresponding to my being in every sense of the term exposed to the world. An embodied agent is one who is in the world as an object and interacts with the world through this "objectivity." In a way, this is exactly what it means to be, or have, a body, as opposed to being, say, a purely mathematical object. The body which we are determines how we are in contact with the world and how we communicate with others. This applies not just to computers, but also to any other artificial cognitive system, as well as to all humans and animals. An agent only interacts with the world, and (in the sense that concerns us here) only knows it, if it exists as more than a mathematical object—if it is also all or part of a material system. Embodiment, construed as a research project, amounts to an exploration of the claim that the material realization of a system has important consequences for its cognitive capacities, so that cognitive agents that are embodied in different ways will have different cognitive capacities.

We humans, like many other biological beings, are embodied as individuals. That is, we are so primarily as distinct entities occupying different places in physical space:[10] ones that cannot simultaneously occupy more than one place, that are to some extent bound to a given place, and that can only ever be in one particular place at a given time. This has important consequences for the way we know and are known by others. First, we are, and show up as being, independent loci of both acting per se and its initiation. The individual agent's presence in physical space is both necessary and sufficient for it to act, and where the agent is located in physical space is also where the agent acts. It may be argued that thanks to modern ICT resources—and perhaps, also, not so modern ones, such as letters—we can also act where we are not ourselves located. For example, I can sell stocks in London while sitting here at my desk in Montreal. This is sometimes true for some of us, but not for all; it is so only for those who have access to complex technical and social systems, in the absence of which such action at a distance is impossible. However, none of us can avoid acting where we are. Because our knowledge as individual agents is related to our ability to act, we are immediately interested and concerned by what we know. That is to say, our knowledge of the world involves each one of us individually in a particular way. For example, my knowledge of "where-the-chair-is" is inseparable from, and colored by, my goal of sitting on or avoiding crashing into it. In consequence, distinct individuals inevitably embody different points of views on the world. Where we are concerned, knowledge is necessarily plural and contradictory.

This basic human characteristic—namely, individual embodiment—also plays a fundamental role in our moral life. Firstly, responsibility can only be attributed to human agents because they are individuals who act independently.[11] That is also why it makes sense to encourage them to pursue certain courses of action and to avoid other courses of action. Secondly, the multiplicity of points of view which this form of embodiment of intelligence necessarily generates opens up a space of dialogue. Because, as individuals, we are individually interested in the world and involved in our knowledge of it, we will at times inevitably disagree, and can therefore criticize each other. Furthermore, individually embodied agents can be punished—something that is impossible for an agent not individually interested in the world.

[10] This, of course, is not a definition of what an individual is, but it does seem that two distinct individuals must at least satisfy the basic condition of having different spatial and temporal coordinates. See Alexandre Guay and Thomas Pradeu, eds., *Individuals across the Sciences* (Oxford: OUP, 2015).

[11] This is not a sufficient condition, but it is a necessary one.

Being individuals thus constitutes a certain way of being in the world that is inseparable from our particular ways of knowing it.

Of course, humans can also know the world in a more abstract way—one where they are only indirectly involved in the results of the inquiry. For example, I may investigate the influence of Hobbes upon Spinoza. This more abstract and objective form of knowledge can, nevertheless, also be of central importance to us, as when we try to develop objective and impartial moral knowledge. However, as Thomas Nagel famously argued in *The View from Nowhere*, an inescapable dilemma arises from the tension between this objective knowledge and individual experience.[12] The main difference between Nagel's approach to this issue and my own is that he construes this tension as resulting from the difference between objective and subjective knowledge, while I seek to anchor the individual's point of view in the objectivity of the body, and insist on the plurality of such points of view. Central to my argument is not the subjectivity of my experience, but rather the objectivity of the fact that when I, or anyone, recognizes the face of someone, diagnoses a disease, or even simply determines the time of the next bus departure at the closest stop, where these are all things artificial cognitive systems typically also do, we as human agents are individually involved in what we know, whereas an artificial system performing an equivalent operation is not interested or involved in the knowledge it produces.

Artificial intelligence, in most cases, is not embodied *as* individuals, but rather embodied *in* systems, where it appears in the form of intelligent artificial agents (or data-driven agents). These agents possess neither individuality nor personhood. They are "analytical agents" rather than material entities. An analytical agent is primarily a mathematical object. It has an environment that is made up of data that represent some aspect of the material world, and the agent responds in more or less complex and autonomous ways to changes in that environment. The agent is part of a complex system that will usually connect together many different types of technology: some electronic and others mechanical, some cognitive and others social—for example, trucks and their drivers, or part of the banking system. This heterogeneous system provides the artificial agent with data, and is constituted in such a way that its (numerical) responses can have consequences in the world. The algorithm or section of code that processes data from the world with certain objectives in view is considered an *agent* because we attribute responsibility to it for whatever it is that the relevant

[12] Thomas Nagel, *The View from Nowhere* (Oxford: OUP, 1986).

systems do in the world, whether it be selling underwear, giving directions, allowing the withdrawal of money from ATMs, or performing diagnoses.[13] On many occasions, there will be more than one intelligent agent involved in obtaining the desired result. The agent is *analytical*, because without the complete system of which it is a part it would not be able to do anything at all—it would not be an agent. In that sense, it is not a real agent, for it cannot act without the whole system of which it is part. Unlike an individual agent that can act by itself, it is only the whole system here that can bring about any transformation in the world. It follows that there is no individual object in the world that corresponds to the agent: the agent is immaterial.

In such cases, which are the most frequently occurring ones, artificial intelligence is not embodied *as* something, but embodied *in* something: embodied, that is, in a system that it itself is not, or to which it does not correspond, yet without which it would not exist or, at least, could not actually figure in the world. In consequence, artificial agents who are analytically responsible for certain events happening in the world are invisible and radically anonymous, where they are the latter just insofar as they are not individual. Put another way, they cannot be individualized otherwise than as mathematical objects. It does not make any sense to ask who they are or where they are. Furthermore, the calculation of whatever function they determine may be distributed over many different physical places, in the internet or "cloud."

At least three important consequences follow from this strange form of embodiment. Firstly, when dealing with artificial intelligence we have the impression that we are faced with some kind of omnipotent and omnipresent invisible entity over which we have little control. The truth is, however, that we are dealing with numerous and sundry material and social systems devised by various persons, enterprises and administrations for different specific purposes: for example, to sell books, to track our movements, to filter job applications, for targeted advertising, to stop suspected terrorists, to control road circulation, to evaluate school teachers or welfare applicants, for targeted policing or political influencing, and so on. The extent to which we who are targeted by them have control over these systems does not depend so much on artificial intelligence as such, as on how they have been made, what the owners/developers wished to accomplish, and

[13] Note, however, that from a causal point of view, the mechanical apparatus that pushes the money bills out so that you can take them is as important to the success of the operation as the system that verifies your password or the algorithm that "decides" that it is the rightful owner of the card who is making the request.

how much room has been left for users to respond or simply opt out. All these technical issues actually reflect the social, commercial and/or political goals of those who adopt and promote these systems. They often also bear witness to their shortsightedness or their misconceptions concerning AI. However, these technical characteristics of the systems will invariably entrench or transform the power relations obtaining between users on the one hand and these dominant social agents on the other.

The second consequence to be noted is that unlike individually embodied agents, who inevitably provide a multiplicity of different and often contradictory points of view, an analytical agent that can deal with thousands or even millions of demands, faces, applicants or diagnoses will impose on all of them just one uniform point of view. There may be some circumstances where such homogeneity in the response constitutes an advantage, but that is certainly not always the case. Especially when the role of the system in which the artificial agent is central is to reject or dismiss people, or to constrain individuals' freedom, this absence of any "minority report" can have dramatic consequences. Interestingly, in Philip Dick's famous short story *The Minority Report*, there are three different individuals—mutants that can predict the future—who do not always agree in their predictions.[14] That plurality is essentially why there can sometimes be a minority report, but it is not the case here. The unicity of the point of view also limits how and what the artificial agent can learn. It can learn through reinforcement to do better what it already does, but in the social domain this improvement in its performance, as has been shown, often corresponds to a self-fulfilling prophecy that makes the agent even more deaf to the claim of those who contest its decisions.[15] What the system cannot learn is how to do something different while doing the same thing. What it cannot gain is a different point of view on whatever it is that it is doing. Here it is worth noting that although this is a technical limitation to the systems we are currently engaged in building, there is little basis for thinking that it is an absolute one affecting all artificial cognitive systems. The simple fact is that we do not want to have automated systems that can "change their mind."

Finally, invisible and immaterial analytical agents that cannot be individualized cannot and should not be held morally or legally responsible for anything. The systems of which they are part and which have consequences in the world are tools created by different persons to achieve specific goals. They are adopted, and implemented, by the same or by other

[14] Phillip K. Dick, *The Minority Report* (London: Gollancz, 2002).

[15] See Eubanks, *Automating Inequality*, and O'Neil, *Weapons of Math Destruction.*

persons for the sake of what they can do: for example, because employers wish to speed up their job applications selection process, or think that AI will make it more objective, or because the owners of trucking companies would like to reduce costs in the form of drivers' salaries and social benefits while also speeding up delivery times by avoiding limits on the number of hours of non-stop driving permitted, where all of this may be accomplishable through introducing self-driven vehicles. Therefore, if harm comes to others because of these tools, clearly it is not the tools themselves that should be held responsible, but rather, as is usually the case with any potentially dangerous technology, those who have decided to use them, who at least in terms of civil law remain accountable even when the damage caused has come about through no fault of their own. In fact, such systems cannot be held responsible because, as was indicated earlier, they are not individually interested in or concerned by the result of their own acting in the world. This as we shall now see, is closely related to our next feature: namely, autonomy.

Autonomy

What, it may be asked, if those artificial agents were autonomous? An analytical agent appears to be but a clog in a complex system, and no matter how sophisticated it may be, it is constrained by the objective pursued by the system as a whole. Its autonomy, then, however complex it may be, seems highly limited. However, what about a self-driving car that has to choose between harming a pedestrian or its passengers, or a military drone that autonomously recognizes its target, evaluates collateral damage, and decides whether or not to fire? Are these machines sufficiently intelligent and autonomous to be held morally or legally responsible for what they do (or don't do)? And are they not embodied differently than analytical agents?

The fact is, that like intelligence, autonomy is a rather ill-defined term and moral autonomy even more so.[16] In robotics and artificial systems, autonomy is usually defined as the ability of a system to adapt by itself to changes in its surroundings. This entails that autonomy is never absolute. It is a relational property, relative to a given environment. Thus, a tiger that is autonomous in the jungle is not autonomous in a zoo, where it cannot satisfy its basic needs, and depends on its human keepers to survive. A robotic lawnmower, for its part, is only autonomous in an environment where there is an electric grid and a station where it can recharge its battery,

[16] Paul Dumouchel, "Philosophy and the Politics of Moral Machines," *Journal of AI Humanities* (forthcoming).

and where the topography of the terrain is such that it can easily circulate between the grass and the recharging station. In the first case, the tiger can lose its autonomy when its environment artificially becomes excessively restricted. The jungle is replaced by some cage in a city zoo. In the second case, if we "open up" the environment, so to speak, the lawnmower appears less autonomous. It is seen to depend on some quite special manmade conditions that allow it to function properly. It is only in this artificially created surroundings that the robotic lawnmower can appear to be, and actually be, autonomous.

This is not entirely surprising. It reflects the fact that, unlike what is the case for a natural system, what it means for an artificial system to adapt to its environment is defined by its maker or programmer. While for a tiger to adapt means to survive and to reproduce, for a robotic lawnmower it means to mow the grass and, when necessary, to go by itself to recharge its battery at the appropriate station. An artificial system that "adapts" to its environment in that predetermined way acts autonomously. The domain of autonomy, or environment, of an artificial agent is always heterogeneously defined with a view to the task it is supposed to accomplish. Whether it be an autonomous car, a military drone, or an algorithm deciding simple cases in lower courts of law, such artificial agents will exert their autonomy in a domain that has previously been determined by their programmer or designer, who has also defined what it is for them to act adaptively. This domain is inevitably narrower and poorer than the environment in which we ourselves live, because for both technical and methodological reasons designers only include into the machine's environment those elements necessary for the system to accomplish the task for which it is designed. The system is simply blind to everything else: for it, those other aspects of the world do not exist.

The autonomy of a natural system, being an expression of the way natural agents are individually embodied, is inseparable from the latter. It reflects the fact that the agent is directly concerned or interested by the consequences of its actions in the world. Its autonomy is a means to the satisfaction of its own basic needs—something which is not the case with artificial agents. It is we, the designers and users who are interested in the consequences of the artificial agent's actions, who define what it is for an artificial system to act autonomously. Unlike a naturally autonomous system, an autonomous artificial agent is not interested as an individual in the consequences of its actions: it is we who care about these. The system is not itself concerned about the consequences of its actions in the world, and *nor do we want it to be so.* That is why it does not make much sense to think of artificial agents as having rights. Rights, as Joseph Raz argued, are essentially based on a person's needs. A right corresponds to a need that is so important and central to an agent that it becomes legally entrenched and morally foundational.[17] However, artificial agents do not have any needs that are proper to them as individual agents. They only have those which we determine are useful for them in relation to their accomplishing of whatever task we ourselves designed and defined them for. That is why, appearances notwithstanding, autonomous vehicles are not individually embodied, but rather, essentially, just analytical agents.

As L. Damiano and I have argued elsewhere, this is not an inevitable limitation where artificial systems are concerned. Rather, it reflects our own choices and objectives as designers of these systems.[18] We do not want to create truly autonomous artificial agents, because we fear what the latter may do: in the main, we do not want robots and artificial agents that will do as they themselves think fit. We are not interested in artificial agents that will, like human workers, criticize us, or refuse to do what they are asked to do because they judge it to be morally objectionable or inappropriate to the task at hand, or simply feel they have done enough for today. We want artificial agents that will do what they are told, and were designed, to do. We want mechanical slaves, not morally autonomous artificial agents. We want artificial agents whose domain of autonomy is very well and precisely defined, where this rules out the possibility of their being morally autonomous.

There are many different definitions of moral autonomy; however, one of its central characteristics is that the domain over which it ranges cannot be precisely defined. It is open-ended. It follows that we cannot rule out *a priori* the possibility of our having, in the future, moral obligations towards artificial agents, and of them even becoming genuine moral agents themselves. Yet for this to happen, they would need to meet some requirements which the artificial systems which we currently interact with on a daily basis do not satisfy. Firstly, they would have to be interested in the world, and in the consequences of their actions, in a specific way: the world and their actions would need to concern them individually—which is another way of saying that being individually embodied, they would have to be *ends in themselves* rather than merely means to an end. This is a minimal condition that is necessary, but probably not sufficient, for

[17] Joseph Raz, *The Morality of Freedom* (Oxford: OUP, 1988).

[18] Paul Dumouchel and Luisa Damiano, *Living with Robots* (Harvard MA: Harvard University Press, 2017).

defining them as moral patients—though certainly not as moral agents. We have moral obligations towards moral patients, even though they do not have any towards us. Because they would be ends in themselves, such artificial agents—just like a dog or an endangered species—would have to be taken into account for the sake of themselves, and given at least some weight when we ask "if the maxim of our action can be transformed into a law of nature.

A morally autonomous agent is one that submits to obligations that, unlike a law of nature, bind the agent, but do not determine its actions. A moral obligation differs from a law of nature in that it does not necessarily bring about a certain state of affairs. This is implicit in Kant's distinction between acting morally and acting in conformity with the moral law. At this point, all moral machines and ethical robots have been designed or imagined as being programmed to act in conformity with the moral law, usually understood as some version of utilitarianism—but there are none that can act morally. Acting ethically, for such machines, is to follow, i.e. to be enslaved to, the (ethical) rules which they are programmed to obey.[19] However, a person or an artificial agent only acts in a morally autonomous way if that agent could have acted otherwise—that is to say, if it could have acted immorally instead.[20] This is what present-day moral machines cannot do; they must follow the moral rules they have been programmed to respect.

Finally—and this brings us to our last feature, which is judgment—a morally autonomous person is one who can recognize that another (be it a person or, perhaps, an artificial agent) has a legitimate claim, even in the absence of any pre-existing moral norm that justifies or grounds that claim. This idea is to be found, for example, in Sen's notion of being "against injustice," which argues that we can recognize injustice in the absence of any theory which defines a given state of affairs as unjust.[21] It is also at the heart of Bergson's concept of "open morality" as opposed to "closed morality." A morally autonomous agent in this sense will be one who is not enclosed within the rules of existing morality: one who does not think that all there is to morality is the satisfaction of a finite set of rules, and who, like the good Samaritan, is able instead to recognize that others can have legitimate claims even when these go against existing rules.

[19] Ibid., 170–95.

[20] Note that a similar requirement applies to law, as Mireille Hildebrandt reminds us. What distinguishes law from technological normativity is that it can be resisted. See Hildebrandt, *Smart Technologies and the End(s) of Law*, 296.

[21] Amartya Sen, *The Idea of Justice* (Harvard MA: Harvard University Press, 2009).

JUDGMENT

To put the point we have just been making another way, a morally autonomous agent will be one that can *judge* that a situation is unfair, unjust, illegitimate, or morally wrong even in the absence of a rule that defines it as such. It is also one that can *judge* that something is appropriate, wonderful, or what morally needs to be done, even in the absence of any rule or prior example showing that this is the right thing to do. As Alessandro Ferrara reminds us, this ability to judge in the absence of a rule lies at the heart of Kant's conception of aesthetic judgment, and he argues that it also grounds a fundamental form of moral normativity.[22]

To judge, in one sense, is to apply a rule, and this is something that artificial cognitive systems can often do extremely well. But in another closely related sense, it is to be able to know when a rule is applicable and when not. This latter is something artificial systems have a lot more difficulty doing. We need to predetermine the type of cases to which the rule applies, and this will rest, in turn, on another rule. At some point, however, the regress must stop, and we come to a point where there is no rule for how to apply another rule—so we need *judgment.* Therefore, judging in this sense is to some extent always done without a rule: that is to say, not necessarily in the complete absence of any rules, but at least partially outside of them—even if just to determine whether some rule applies to this set of circumstances or not.

Finally, while the two meanings given above both involve realizing that something counts as an instance of a (pre-existing) rule, to judge can also sometimes mean discovering the rule in the example. This last meaning of the term is how Kant conceived of aesthetic judgment, and it corresponds to what Ferrara defines as *exemplary* validity. An exemplary action, or work of art, constitutes an exception or transgression, something that does not fall under existing rules, and which becomes the guiding principle or paradigm of a new rule. It asserts itself, and is recognized for its value, over against (or at least outside of) existing rules. In this sense, actions and decisions of exemplary quality lie at the heart of our moral and cognitive progress and discoveries. Now what interests me, and what I believe to be central in the context of this last meaning of the term, is the contradictory aspect of the cognitive operation involved: to judge in this sense requires one to recognize that something contradicts or falls outside of the rule, and yet not simply discard it as meaningless or a purely negative instance

[22] Alessandro Ferrara, *The Force of the Example. Explorations in the Paradigm of Judgment* (New York, NY: Columbia University Press, 2008).

of something—to recognize it, that is, as meaningful nonetheless, either in itself or relative to the very rule it contradicts. Precedents in common law sometimes play such a role, functioning either as something that exemplifies a rule or as prompting us, in an exemplary way, to transform its meaning and domain of application.

To judge without a rule is something that today's artificial systems are unable to do. This deficiency is partially technical, it being due to the way in which they learn. It is also due to the way they are embodied. The fact that they are not individually embodied, and therefore cannot be morally autonomous, condemns them to only ever instantiate a unique point of view—one which, because of the fact that they learn through reinforcement, they can only reassert in all circumstances. This makes it impossible for them to learn through contradiction. The human being's ability to judge is, I believe, a consequence of what Hannah Arendt called "the human condition of plurality, the fact that men, not Man, live on the earth and inhabit the world."[23] In other words, it is the plurality of points of view which individually embodied morally autonomous agents bring with them that allows us to judge—to recognize—that what falls outside the rule can also be the foundation of a new rule, that its failure to fit the rule does not make it meaningless.

It may, of course, be objected that humans are not very good at judging. Examples of bad judgment abound in both the private and public spheres. This is certainly the case. However, as this objection itself clearly illustrates, humans do distinguish between good and bad judgments, and they are impressed by exemplary good (and bad) judgments, which are exemplary precisely because they give rise to new rules or norms of what should or should not be done. We may disagree with each other about what constitutes a good judgment, but the language of judgment, and of exemplary judgments, is one in which we humans are ourselves conversant. Artificial cognitive systems are not.

Interestingly, contradictory judgments are at the heart of some of the most important intellectual activities human beings engage in: law, politics, science, philosophy and theology. At the same time, in all of these areas there has always existed a tendency towards unification, towards passing over or erasing such disagreements and contradictions and replacing them by a single unique point of view. The central characteristics of present-day artificial intelligence suggest that it constitutes another expression of that

[23] Hannah Arendt, *The Human Condition* (Chicago, IL: University of Chicago Press, 1958), 7.

same age-old tendency, and that it is therefore unlikely to transcend the human condition.

Bibliography

Arendt, Hannah. *The Human Condition.* Chicago, IL: University of Chicago Press, 1958.

Barrat, James, R. *Our Final Invention: Artificial Intelligence and the End of the Human Era.* New York, NY: Dunne/St. Martin, 2015.

Bongard, Josh, and Rolf Pfeifer. *How the Body Shapes the Way We Think.* Cambridge, MA: MIT Press, 2007.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies.* Oxford: OUP, 2014.

Chemero, Anthony. *Radical Embodied Cognitive Science.* Cambridge, MA: MIT Press, 2009.

Dick, Phillip, K. *The Minority Report.* London: Gollancz, 2002.

Dumouchel, Paul. "Philosophy and the Politics of Moral Machines". *Journal of AI Humanities* (forthcoming).

Dumouchel, Paul, and Luisa Damiano. *Living with Robots.* Translated by Malcolm DeBevoise. Cambridge, MA: Harvard University Press, 2017.

Eubanks, Virginia. *Automating Inequality.* New York: St-Martin Press, 2017.

Ferrara, Alessandro. *The Force of the Example. Explorations in the Paradigm of Judgment.* New York: Columbia University Press, 2008.

Ganascia, Jean-Gabriel. *Le mythe de la singularité.* Paris: Seuil, 2017.

Guay, Alexandre, and Thomas Pradeu, eds. *Individuals across the Sciences.* Oxford: Oxford University Press, 2015.

Hildebrandt, Mireille. *Smart Technologies and the End(S) of Law.* London: Edward Elgar, 2015.

Nagel, Thomas. *The View from Nowhere.* Oxford: OUP, 1986.

O'Neil, Cathy. *Weapons of Math Destruction.* New York: Crown, 2016.

Pfeifer, Rolf, and Josh Bongard. *How the Body Shapes the Way We Think.* Cambridge, MA: MIT Press, 2007.

Raz, Joseph. *The Morality of Freedom.* Oxford: Oxford University Press, 1988.

Sen, Amartya. *The Idea of Justice.* Cambridge, MA: Harvard University Press, 2009.

Shen, Guohua, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. "Deep Images Reconstruction from Human Brain Activity." *bioRxiv.* Accessed October 1, 2019. https://doi.org/10.1101/240317

Stewart, James, Olivier Gapenne, and Ezequiel A. Di Paolo. *Enaction. Towards a New Paradigm for Cognitive Science.* Cambridge, MA: MIT Press, 2010.

Vidal, Fernando, and Francisco Ortega. *Being Brains. Making the Cerebral Subject.* New York, NY: Fordham University Press, 2017.