Is My Car Evil? A Review of Non-Anthropocentric Theories of Moral Agency

Tommaso Barbetta[*]

Introduction

January 22, 2018. It is Monday morning and a huge firetruck is stationing in the middle of the Interstate 405, in California. A car, travelling at 100 km/h, proceeds right toward the truck without slowing down. Nothing obstructs the view of the driver. There is no malfunctioning in the electrical system and the hydraulic brakes are fully operative, ready to be activated. Yet, the driver does not push the brake pedal. The car continues its run, getting closer and closer to the firetruck parked in the centre of the road. The impact seems now inevitable. The vehicle crashes into the back of the red truck. The metallic front of the car folds in on itself. Surprisingly, the driver is safe. This time nobody got injured.[1]

More than 1 million people die every year in traffic accidents all over the world (WHO 2015). However, quite curiously, this banal collision caught the attention of some of the most popular newspapers around the globe. Why was this relatively harmless incident taken so seriously by the press? One of the reasons is that, apparently, the man sitting at the driver seat was not actually driving the car during the accident. But not because he fell asleep at the wheel, nor because he was distracted by a notification from his smartphone. He deliberately chose not to drive. Indeed, someone else was driving the car in his place. But here is the issue: that someone, the entity that was really behind the wheel during the accident, was not a human being. It was the "autopilot" . It was a piece of software.

Nowadays smart technologies substituting us and acting for us are everywhere. We are quickly getting used to delegating everyday tasks to a multitude of artefacts. Artefacts that constantly mediate our experience of reality and help us in the process of making decisions. Yet, from a moral point of view, we do not really know how to consider all these entities.[2] What can be said about the above-mentioned software that destroyed a car and that could

[*] University of Tokyo, Graduate School of Interdisciplinary Information Studies, ITASIA D2

キーワード：Actor-network theory, Information ethics, posthumanism, moral agency, responsibility, morality of technology.

have killed its owner? Is it responsible for the accident? Is the company that designed it responsible? Or should we rather blame the human driver, who was supposed to - but presumably did not - keep his hands on the steering wheel and be vigilant in order to take control of the car before the collision. And what about the algorithm that in 2012 made Knight Capital lose 440 million USD in 45 minutes?[3] Can it be blamed for the disruption of the company? Nobody designs trading algorithms with the intention of breaking a company apart or causing a market to crash, but that is nonetheless what they ended up doing in some cases (MacKenzie 2014, 3). On the surface, (quasi-)autonomous technologies appear to behave as if they could make decisions and act on their own. However, the lack of intentions - besides those delegated by humans - behind such Boolean logic, has led consequentialists to cast serious doubts over the significance of recognizing the actions performed by artefacts as morally charged (Peterson and Spahn 2011). Due to their devotion to the idea of *intentionality*, standard ethical theories have been unable to recognize the moral implications of information and communication technologies (ICT) such as the one adopted by the above-mentioned autopilot.

Outside the field of ethics, over the last 30 years several non-anthropocentric theories of agency have emerged. In particular, actor-network theory (ANT) has provided an effective framework for the assessment of the role played by artefacts, and other kinds of non-human entities, in our society.[4] However, it is not yet clear if and how a notion of agency detached from intentionality, such as that of ANT, could be translated to ethics. Once it is freed from intentionality, is moral agency a characteristic of a specific category of entities, such as humans, living beings or algorithms, or could it be ascribed to anything? Furthermore, how could the question of responsibility be reframed in order to fit an ontology according to which agency is always distributed among multiple actors?

After a brief overview of the instrumentalist approach to technology embraced by standard anthropocentric moral theories, the present paper investigates these questions by reviewing the literature of ANT (along with its postphenomenological adaptation). Furthermore, in the final sections, the article turns to information ethics (IE), a non-anthropocentric ethical theory influenced by computer science, whose different definition of moral agency might help us to solve some of the difficulties of an ANT-informed ethics.

1 Instrumentalism

From the point of view of standard contemporary ethical theories, such as deontology, contractualism and consequentialism, it would be incorrect to consider an artefact as a moral agent. In standard ethics by definition an action is moral only as long as it is initiated by some kind of intentional state of mind, i.e. only if it is the product of a conscious reason (Himma 2009, Koops et al. 2010). Given this definition, there is no way my car could be a moral agent since, due to the limits of our current technologies, a computer cannot possibly have intentional states of mind.

By accepting intentionality as a necessary condition for moral action, standard ethical theories therefore set a clear separation between human entities, which reflect, make decisions and have intentions and free will, on the one hand, and purely neutral artefacts, on the other. To rephrase the (in)famous slogan of the American National Rifle Association: cars don't kill people, people kill people. An artefact – be it a car or a gun - is only a mere instrument, and as such it has always different possible uses. It is ultimately the human user who determines what to do with it (Pitt 2014). Responsibility lies always in a human.

Because of their commitment to the idea of moral responsibility, standard ethical theories are incapable of recognizing many of the pressing moral issues of our information society. First of all, a strong anthropocentric approach seems to ignore the existence of artefacts that are able to learn from the environment, correct themselves and make decisions that were not intended either by their producers, or by their users. Take the case of the autopilot mentioned in the introduction, which relies on a computer vision system: this system was built on an artificial neural network where code is not entered line by line by human programmers and whose output is often obscure to the original creators themselves. It would be inaccurate to hold a human programmer responsible for the individual choices made by such kinds of software (Matthias 2004).

Moreover, standard ethical theories seem to downplay the role of technologies in mediating the state of mind of individual humans. The increasing adoption of neuromarketing strategies attests the widespread awareness among commercial corporations of the influence that ad-hoc physiological stimuli might have on the behaviour of consumers. For example, the layout and atmospherics of most casinos are designed to make customers move following specific patterns and keep them playing as long as possible (Schüll 2012). This kind of artificial environment is by no means neutral. Rather, it actively manipulates the perceptions of its users. What we call addiction is not the unidirectional product of the harmful decisions made by an (genetically) impaired brain. It is a relational phenomenon involving a human and an environment often - but not necessarily - designed to make use of his limitations.

The following sections investigate advantages and disadvantages of the adoption of a notion of agency influenced by ANT, focusing on the moral issues raised by the crash.

2 Actor-network theory

Developed in the early 1980s in the field of sociology of scientific knowledge, ANT is a heterogeneous set of linguistic categories that initially emerged to trace the active role of non-human entities in the analysis of scientific practices (Latour 1983). The common ground, shared by the social scientists adopting this analytical framework, is a non-anthropocentric view of reality based on the concept of network. According to Bruno Latour, the author who more than anyone else has unfolded the philosophical consequences of ANT, the best way to take into consideration how a heterogeneous set of non-human entities, from microbes to airbags, participates in our collective life is to adopt a “network-like ontology” (Latour 1996, 370).

The next sections investigate the ethical implications of this ontological stance. Following a review of the key concepts of symmetry and mediation, the article discusses how the notions of “prescription” and “distributed agency” might inform a moral analysis of non-human entities.

2.1 Principle of symmetry

ANT is based on the idea that there is an ontological symmetry between humans and non-human entities. This is sometimes referred to as the “principle of generalized symmetry” (Callon 1986). Human beings, biological organisms, material objects and abstract entities are all part of the same reality. Given this principle, a common terminology, one which is not biased by the previously mentioned human vs. non-human dichotomy, is needed by social scientists. “Actor”(or “actant”) is the term used to express the minimum unit of our reality. An actor is literally anything that acts, any node that contributes to a network. An actor is defined by its action, and since “there is no other way to define an action but by asking what other actors are modified, transformed, perturbed or created” (Latour 1999, 122), it follows that an actor is necessarily defined by its relations with other actors within a network.

According to John Law (1992), from an analytical standpoint this symmetrical approach leads us to a radical rejection of any difference in kind between humans and objects. Humans are not at the centre of the universe anymore.

They are equal to any other entity. However, it would be a mistake to read ANT's analytical norms as moral norms. Indeed, justifying a moral symmetry on the basis of an ontological symmetry would be logically flawed: the fact that reality is in a certain way does not imply that we ought to act accordingly - something which was pointed out already by Hume (1896). "We need, I think, to distinguish between ethics and sociology. The one may - indeed should - inform the other, but they are not identical. To say that there is no fundamental difference between people and objects is an analytical stance, not an ethical position. And to say this does not mean that we have to treat the people in our lives as machines." (Law 1992, 383). Law is – rightly – afraid of the consequences of a misinterpretation of the principle of symmetry. Yet, he leaves the door open for a possible contribution of ANT to ethics.

In an article titled "Morality and Technology" , Latour (2002) takes on this task and attempts to elaborate an ANT-informed view on ethics. In the article, Latour openly criticizes standard ethical approaches that divide technology and morality into two separated realms, that of means, on the one hand, and that of ends, on the other. Artefacts cannot be reduced to mere means, since they lay the conditions for our actions, and thus mediate our behaviour. According to Latour, what we call "human" cannot exist independently from the technological mediations it is intertwined with (Latour 2002, 252). "Generalizing the notion of affordance, we could say that the quasi-subjects which we all are, become such thanks to the quasi-objects which populate our universe" (Latour 2002, 252-253). We, as quasi-subjects, are free either to accept or reject the programmes of action - the affordances - embodied by the artefacts surrounding us, but nonetheless our behaviour is necessarily mediated by them. A moral approach informed by the relational ontology of ANT, is therefore one that focuses on technical mediations.

2.2 Mediations

In ANT the way actors interact with each other is referred to as mediation. Two actors, two nodes of a network, always interact with each other by the means of a third actor (Latour 1996, 378). Given an actor A and an actor B, their interaction necessarily occurs through the mediation of an actor C. This third actor C is a "mediator" (Latour 1996, 373). In the case of non-autonomous cars, for example, human drivers control the wheels using the steering wheel. At this level of abstraction, the steering wheel is a mediator. However, the steering wheel is merely an interface and there is a series of other mediators between it and the wheels: e.g. a steering column, several sensors, an electronic control unit and a motor, all translate the mechanical torque of the steering wheel operated by the driver into the actual steering. Moreover, it has to be remarked that this is not a one-dimensional chain, but a heterogeneous network. By assuming different levels of abstraction, we would recognize that different kinds of entities mediate the way we control a car: roads and traffic signals obviously play a part, but also the voice of a car navigation system telling us what to do, as well as the interiorized gaze of the authority that keeps us from breaking traffic rules. Modifying any of these actors might produce a different driving experience and different subjectivities.

The autopilot illustrates how the introduction of a new mediator might change the way humans perceive the environment and act. The control of the car is delegated to software, which dissociates the body of the human driver from the movement of the vehicle. The car keeps driving on its own. However, the self-driving systems currently available have not yet reached full autonomy and from time to time require the human driver to take control of the vehicle. The human driver is now supposed to assume the role of "drive monitor", ready to take control of the car whenever the software is perceived to be doing something wrong. Given this new role, it seems that in our example both the human and the software made a mistake. The human was wrong since he did not correct the autopilot. Yet, it has been demonstrated that the high level of automation of current software lets humans disengage from the driving task, to such an extent that they become potentially less attentive and therefore unable to take control when necessary (Banks et al 2018). Until recently, driving has been a bodily experience: drivers received a feedback at every movement, and contextual cues activated habitual responses. The autopilot does not just change the car; it also changes the human driver as a subject. The way we look and feel the road is different.

2.3 Prescription

Initially employed by Madeleine Akrich (1992), the notion of prescription conceptualizes one of the key insights of ANT: i.e. the idea that humans are able to inscribe programs of action into things and act at distance. In the process of creation of an artefact, designers envision the way the artefact is assumed to interact with us and inscribe such vision in it. "Like a film script, technical objects define a framework of action together with the actors and the space in which they are supposed to act" (Akrich 1992, 208).[5] Software is the most obvious example of prescription - it literally is a collection of scripts encoded by a programmer -, but the concept can be extended to any kind artefact even outside the realm of ICT. Latour gives the example of the speed bump (Latour 1999, 185), a very simple technology used to make drivers slow down. In this case the script is materially built into the road.

The notion of prescription has several ethical implications. Scripts might be implemented to produce “moral delegations” (Latour 1999, 217). Designers can inscribe moral instructions into technologies, in order to make users behave accordingly. This commonly occurs through patterns of punishment and rewards. The acoustic signals and visual messages produced by a car, telling the driver to fasten his seatbelt - or, in more recent cars, to hold the steering wheel while a self-driving system is enabled -, are examples of moral delegation.

Building on this notion, the postphenomenologist Peter Paul Verbeek stresses the necessity for a proactive approach to technology. Given “that technologies inevitably play a mediating role in the actions of users” , what we need to do is to moralize our technological environment (Verbeek 2006, 377). But who is supposed to moralize our technologies? According to Verbeek, the script approach “reveals a specific responsibility of the designer, who can be seen as the inscriber of scripts” (Verbeek 2006, 362). It is thus the duty of designers to anticipate technological mediations and moralize technology in accordance.

Design becomes a political matter and has fundamental moral implications. However, we should not forget that designers do not act autonomously. They are part of larger networks. Designers typically operate inside organizations and follow decisions taken elsewhere – e.g. the product planning department. Moreover, the notion of “good” is context-dependent and might diverge fundamentally depending on the aims of the organization in which designers operate. From the point of view of a car maker, the best autopilot is not necessarily the safest one. Indeed, at the current technological stage, the safest autopilot would probably be one that needs the human driver to never release the steering wheel and be constantly attentive. However, this would undermine the convenience of the system itself, making it less appealing to consumers, something which the CEO of the company that manufactured the car mentioned in the introduction clearly recognizes: “This is crux of matter: can’t make system too annoying or people won’t use it” .[6] A few months after the accident the same company released a software update that strongly increased the frequency of the “hold the steering wheel” alert signals whenever a driver is not touching the steering wheel.[7] As far as we know, a different prescription might have made the driver take control of the car in time, thus avoiding the accident. Yet, blaming the designers for letting the driver get too complacent would be inaccurate, since it neglects the larger network that determined such design in the first place and erases the agency of the human behind the wheel.

2.4 Distributed agency

While the concept of mediation is used to deconstruct the interaction between two actors, the idea of distributed agency (Callon and Muniesa 2005) allows us to analyse an event, such as the incident described in the introduction, as a network composed by a sum of microtransactions.[8] The concept in itself does not define a special kind of agency: in principle, any action can be seen as distributed across a network.

Jane Bennet's analysis of the North American blackout that occurred in 2003 is a clear example of how this concept can be applied (Bennet 2005). According to Bennet, the blackout that affected 45 million people cannot be reduced to one individual cause and no singular entity can be held responsible for it. Human omissions, the growing demand for electricity by a collective of consumers, legal deregulations, economic transactions, the movement of electrons, etc. - the effects of all these actions put together have collectively contributed to the blackout. Once they intersect, a broad set of small - apparently negligible - actions performed by different kinds of entities, might generate huge consequences.

While the analytical advantages of the notion of distributed agency are fairly clear, the moral implications are more complicated. The idea of responsibility is at stake here. In the case of the crash, claiming that it is the assemblage composed by the driver plus the sensors plus the software plus the engineers plus the stationary firetruck and so on, that is responsible for the collision of the car ends up emptying the notion of responsibility of its meaning, transforming it into a useless category. It seems that ultimately nobody should be recognized as responsible: neither the manufacturing company, nor the engineers that developed the software, nor the owner of the car, nor the law, nor the software itself. Bennet goes as far as to claim that the idea of "strong responsibility" is empirically false (Bennet 2005, 463). However, even this assumption does not necessarily imply an incompatibility between the concept of distributed agency and moral analysis. "A distributive notion of agency does interfere with the project of blaming, but it does not thereby abandon the project of identifying (what Arendt called) the sources of harmful effects" (Bennet 2005, 463). Bennet is quietly suggesting that moral analysis does not necessarily coincide with responsibility-assignment. An idea that – as we will see - is also at the centre of Floridi's work on information ethics.

2.5 Ethical issues of ANT

As Verbeek has stressed in his work, the idea that technological mediations shape our perception of reality and frame our behaviours is a major contribution of ANT to the field of ethics (Verbeek 2005, 2006). According to Verbeek, this insight encourages a proactive approach to ethics: given that purely autonomous humans do not exist and that our moral decisions are necessarily mediated by technologies, the new task of ethics is that of moralizing our material environments. But what moral orientation should designers follow in this quest for moralizing technologies? What actors should they consider as the moral patients of their work? Humans, biological organisms, organizations? ANT cannot tell us what is good and what is wrong, or how we are supposed to act in specific situations. This is because ANT does not assess which entities should be considered as moral patients - i.e. the objects of moral action. This is not necessarily a limit, since ANT does not aim at producing a normative theory of ethics. Rather, it is supposed to provide an agnostic tool-kit for ethical analysis, which, in principle, could be adopted by different ethical theories. According to ANT's flat ontology, any theory of ethical patientness would in fact be equally contingent.

The limit of ANT is not so much the lack of a definition of moral patients, but rather the lack of a clear definition of the category of moral agents. An army of drones and a tsunami are both actors that could kill people, but are they both also moral agents? Following Verbeek's phenomenology, it could be argued that ultimately the discriminant between these two forms of agency is human intentionality. From this point of view, things behave morally whenever they are human products: drones are moral agents, while a tsunami is not. Such a definition, however, seems to contradict the principle of symmetry by assuming an a priori distinction between human and non-human actors. Due to this contradiction, this definition partially falls back into a weak form of anthropocentrism and does not offer a persuasive explanation for the behaviour of artefacts that do not follow the script of their creators.

In order to avoid the ghosts of anthropocentrism a different definition of moral agency is needed. The next sections explore how IE has tackled this same issue and has produced a coherent definition of moral agency.

3 Information Ethics

Information ethics (IE) has its origins in the late 1990s, when Luciano Floridi proposed it as the theoretical counterpart of computer ethics – a subject that at that time was largely neglected by moral philosophers (Floridi 1999). IE has emerged within the broader field of philosophy of information. Developed in the context of the expansion of ICT, philosophy of information has two fundamental aims according to Floridi: clarifying the nature of what we call information, on the one hand, and investigating the possible philosophical applications of frameworks and methodologies developed in the field of computer science, on the other (Floridi 2011a, 14).

Like ANT, IE is also informed by a non-anthropocentric understanding of reality. However, their ontologies are based on entirely different assumptions. ANT is a relational ontology. Being necessarily means "being in relation" to something. IE, instead, is based on an informational ontology, and claims that information is the lowest common denominator shared by any entity (Floridi 2010, 94). This ontological assumption leads Floridi to define IE as an extension of ecological ethics: "all entities, qua informational objects, have an intrinsic moral value" and therefore have to be taken into account as moral patients (Floridi 2010, 89).[9] Thus, in contrast to ANT, IE provides a clear definition of the category of "moral patients" . Adopting the concept of *infosphere* - the informational adaptation of the idea of the biosphere - Floridi goes as far as to claim that, since every informational entity is a moral patient, the general moral principle according to which any action should be oriented to is that of avoiding entropy - i.e. loss of information.

3.1 Moral agents

Adopting the method of levels of abstraction (LoA)[10], IE draws a distinction between the category of "moral patients" and that of "moral agents" . As already mentioned, IE considers every informational entity populating the infosphere as a moral patient. However, it specifies that not every informational entity is necessarily also a moral agent. Moral agents are a subclass of the larger category of moral patients. In contrast to ANT, IE clearly defines what kind of actions and what kind of entities can be morally qualifiable. According to Floridi and Sanders (2004), to qualify as a moral agent, an entity must be:

1. Interactive: it has an input and an output, through which it interacts with the environment.
2. Autonomous: it has relative control over its internal condition, so that it can perform an action without the direct command of other actors.
3. Adaptable: it can learn - i.e. it can change the internal rules that determine its actions, in response to interactions with the environment.

Before examining the implications of these criteria, it is necessary to clarify that the definition of autonomy presented here does not coincide with that repudiated by ANT. Floridi's definition of autonomy does not imply the possibility of pure autonomous decisions, nor does it imply the existence of some kind of transcendental subjectivity or free will. It is a quasi-autonomy: a partial control that a system can exert on its internal state. Moreover, depending on the LoA that we choose, a (quasi-) autonomous system can be decomposed into a network. For example, "depending on the LoA adopted, the autopilot can be considered as a single actor that performs the operation of flying an airplane or as a set of interacting actors that execute the subtasks of that operation" (Turilli 2011, 377).

Let us now examine the implications of such a definition of moral agency. First of all, in IE moral agency is not limited to individual humans, but can be attributed to biological organisms, to organizations, and to IT-artefacts, provided that they follow the three criteria just mentioned. However, actions of entities that do not meet these criteria, such as a tsunami or a speed bump, cannot be accounted as moral. Can an autopilot software be considered a moral agent from this perspective? It depends. If it is capable of making decisions and learning from the environment - e.g. using reinforcement learning -, yes. However, what if the car is not able to autonomously change its internal rules, but collects data that is then used to train the software through supervised-learning? Even though the car would ultimately be able to change its internal rules by downloading and updating the software, and even though the code would be written mostly by machines rather than human programmers, in this case the individual vehicle would not be considered a moral agent. However, we can adopt a higher LoA, and look at the car as part of a larger system that includes the neural networks adopted, as well as the engineers that tweak the software, monitor its learning process and release updates. At this LoA, the car could be considered as part of a larger moral agent.

A second implication is that intentionality is not accounted as a necessary condition of moral agency. This leads us back to Bennet's comment concerning "strong responsibility" (see 2.4). Bennet claimed that, if agency is distributed across a network, it is not possible to appoint individual actors as morally responsible for an event. A notion of moral agency without intentionality encounters the same obstacle. An autopilot might be able to take autonomous decisions, but can it be blamed for its mistakes? Even if that was the case, due to the lack of self-consciousness, attributing legal personhood to IT-artefacts and punishing them in case of wrongdoing seems completely meaningless - if not impossible (Koops et al. 2010). This dilemma makes us face what Matthias has called the "responsibility gap" (Matthias 2004), a condition of increasing distance between human creators/users and IT-artefacts, where nobody is liable for the wrongdoing of a machine.

Floridi is able to solve this apparent deadlock by drawing a distinction “between *moral responsibility*, which requires intentions, consciousness and other mental attitudes, and *moral accountability*” (Floridi 2011a, 88). A distinction, which, according to the author, finally frees normative ethical theory from the shadow of anthropocentrism and, most importantly, from the “regress of looking for the responsible individual when something evil happens” (Floridi 2011a, 88). However, while arguing that responsibility-oriented ethics has been unable to acknowledge the role of artificial agents, Floridi does not dismiss the concept of responsibility in toto. In IE humans have responsibilities towards the whole infosphere. They bear “ecopoietic responsibilities” (Floridi 2011a, 91) - i.e. they are responsible for the creation and the well-being of the environment. Similarly to what is suggested by ANT, humans have a peculiar position in the moral outlook of IE: they are not the only moral agents in this world, but they are nonetheless special due to their ability to create artefacts.

Conclusion

Focusing on the example of a self-driving car, this article has reviewed the possible advantages, as well as the limits, of an ANT-informed theory of ethics, and has briefly illustrated the alternative definition of moral agency provided by IE. Despite their different ontological foundations, ANT and IE have encountered similar challenges in the development of a non-anthropocentric moral approach and have often reached comparable conclusions in tackling some of these issues.

ANT lets us recognize artefacts as moral agents and gives us the tools to deconstruct any event into a network. It allows us to investigate how artefacts concretely mediate our ethical decisions - e.g. in terms of prescriptions. However, it also runs the risk of falling into a bottomless relativism according to which nothing/nobody can ever be blamed. IE seems to avoid this relativist deadlock by providing a narrower definition of moral agency, which focuses exclusively on autonomous entities capable of learning from the environment and changing their internal rules. Ultimately, both ANT and IE attempt to shift the focus of ethical analysis from *moral responsibility* to *moral accountability*. The two theories would argue that, while a self-driving car cannot be responsible for an accident, it could, nonetheless, be a source of harmful effects: i.e. it could be seen as morally accountable. However, both ANT and IE have not been clear enough in distinguishing and defining these two concepts. How does “weak” responsibility differ from moral accountability, and do these two concepts imply the existence of two different kinds of moral agency - e.g. human vs. non-human moral agency? These are issues that need further investigation.

[1] Peter Valdes-Dapena, 2018, “Tesla in Autopilot mode crashes into fire truck.” *CNN*, http://money.cnn.com/2018/01/23/technology/tesla-fire-truck-crash/index.html, accessed 01/09/2018.

[2] In this text the terms morality and ethics are used interchangeably.

[3] Matthew Philipps, 2012, “Knight Shows How to Lose $440 Million in 30 Minutes” , *Bloomberg*. https://www.bloomberg.com/news/articles/2012-08-02/knight-shows-how-to-lose-440-million-in-30-minutes, accessed 01/09/2018.

[4] The focus on ANT and Information Ethics (IE) is motivated by their explicit reference to non-human moral agency. Due to the lack of space this paper does not consider the moral implications of other posthumanist approaches.

[5] Is should be stressed that in contrast to technological determinism, scripts can be more or less flexible. Users of technologies might resist against scripts by rejecting or by hacking them (Oudshoorn et al. 2002).

[6] Elon Musk, 2018, *Twitter*. https://twitter.com/elonmusk/status/1005879049493725186, accessed 01/09/2018.

[7] Fred Lambert, 2018, *Electrek*, https://electrek.co/2018/06/11/tesla-autopilot-update-nag-hands-wheel/, accessed 01/09/2018.

[8] A similar idea can be found in the work of several authors influenced by ANT. Latour (1999) uses the term “composition” , while Bennet (2005) talks about “agency of assemblages” and “distributive agency” .

[9] According to Doyle (2010), it is unclear why informational entities would have an intrinsic moral value.

[10] For a detailed explanation of the method of levels of abstraction see Floridi (2011b).

Bibliography

Akrich, Madeleine. "The de-scription of technical objects." In *Shaping Technology/Building Society: Studies in Sociotechnical Change*, edited by Bijker and Law, 205-224, 1992.

Banks, Victoria, and A. Eriksson, J. O'donoghue, N. Stanton. “Is partially automated driving a bad idea? Observations from an on-road study.” *Applied ergonomics* 68 (2018): 138-145.

Bennet, Jane. “The Agency of Assemblages and the North American Blackout.” *Public Culture* 17, no. 3 (2005): 445–65

Callon, Michel. "Some elements of a sociology of translation." *Power, action and belief: a new sociology of knowledge*? London, Routledge, 196-223, 1986.

Callon, Michel, and Fabian Muniesa. “Peripheral Vision.” *Organization Studies* 26, no. 8 (2005): 1229–1250.

Doyle, Tony. "A Critique of Information Ethics." *Knowledge, Technology & Policy* 23, no. 1-2 (2010): 163-75.

Floridi, Luciano. "Information ethics: On the philosophical foundation of computer ethics." *Ethics and information technology* 1, no. 1 (1999): 33-52.

Floridi, Luciano, and J.w. Sanders. "On the Morality of Artificial Agents." *Minds and Machines* 14, no. 3 (2004): 349-79.

Floridi, Luciano, ed. *The Cambridge handbook of information and computer ethics*. Cambridge University Press, 2010.

Floridi, Luciano. *The philosophy of information*. Oxford University Press, 2011a.

Floridi, Luciano. "The method of levels of abstraction." *The Philosophy of Information*, 2011b.

Floridi, Luciano. "Distributed morality in an information society." *Science and engineering ethics* 19, no. 3 (2013): 727-743.

Himma, Kenneth Einar. "Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?" *Ethics and Information Technology* 11, no. 1 (2009): 19-29.

Hume, David. *A Treatise of Human Nature*. Oxford: Clarendon Press, 1896.

Koops, Bert-Jaap, Mireille Hildebrandt, and David-Olivier Jaquet-Chiffelle. "Bridging the accountability gap: Rights for new entities in the information society." *Minn. JL Sci. & Tech.*11 (2010): 497.

Latour, Bruno. "Give me a laboratory and I will raise the world." *Science observed* (1983): 141-170.

Latour, Bruno. "Where Are the Missing Masses? The Sociology of a Few Mundane Artefacts" (1992).

Latour, Bruno. "On actor-network theory: A few clarifications." *Soziale welt* (1996): 369-381.

Latour, Bruno. *Pandora's hope: essays on the reality of science studies*. Harvard university press, 1999.

Latour, Bruno. "Morality and technology." *Theory, culture & society* 19, (2002): 247-260.

Law, John. "Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity." *Systems practice* 5, no. 4 (1992): 379-393.

MacKenzie, Donald. "A sociology of algorithms: High-frequency trading and the shaping of markets." *Unpublished paper* (2014).

Matthias, Andreas. "The responsibility gap: Ascribing responsibility for the actions of learning automata." *Ethics and information technology* 6, no. 3 (2004): 175-183.

Oudshoorn, Nelly, Ann Rudinow Saetnan, and Merete Lie. "On gender and things: Reflections on an exhibition on gendered artifacts." In *Women's Studies International Forum*, vol. 25, no. 4, pp. 471-483. Pergamon, 2002.

Peterson, Martin, and Andreas Spahn. "Can technological artefacts be moral agents?" *Science and Engineering Ethics* 17, no. 3 (2011): 411-424.

Pitt, Joseph C. " “Guns Don’t Kill, People Kill” ; Values in and/or Around Technologies." In *The moral status of technical artefacts*, pp. 89-101. Springer, Dordrecht, 2014.

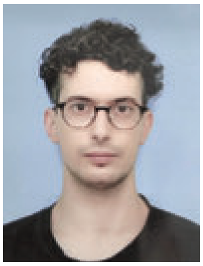
Schüll, Natasha Dow. *Addiction by design: Machine gambling in Las Vegas*. Princeton University Press, 2012.

Turilli, Matteo. “Ethical Protocols Design” , in *Machine ethics*. Edited by Anderson, Michael, and Susan Leigh Anderson, 375-397, Cambridge University Press, 2011.

Verbeek, Peter-Paul. *What things do: Philosophical reflections on technology, agency, and design*. Penn State Press, 2005.

Verbeek, Peter-Paul. "Moralizing Technology: on the morality of technical artifacts and their design." In *paper for workshop, Utrecht University (24 pp)*. 2006.

World Health Organization. “Global status report on road safety 2015” .



Tommaso　Barbetta（とまぞ・ばるべった）

［生年月］1990年5月
［出身大学または最終学歴］ヴェネツィア・カ・フォスカリ大学　アジア・北アフリカ研究家修士課程修了
［専攻領域］STS, プラットフォーム・スターディーズ
［所属］東京大学大学院学際情報学府博士課程

Is My Car Evil?
A Review of Non-Anthropocentric Theories of Moral Agency

Tommaso Barbetta*

Smart-technologies substituting us and acting for us have become increasingly ubiquitous over the last decade. We are quickly getting used to delegating everyday tasks to a multitude of artefacts. Artefacts that constantly mediate our experience of reality and help us in the process of making decisions. However, from a moral point of view, we do not yet know how to consider all these entities.

In the social sciences, Actor-network Theory (ANT) has provided a consistent framework for the analysis of non-human agency. This has been theoretically possible thanks to the detachment of the notion of agency from that of human intentionality. However, it is not clear if and how a notion of agency detached from intentionality could also be embraced by the field of ethics. What is the usefulness of ascribing moral agency to non-human entities? Would such a new notion of moral agency be a characteristic of one specific category of entities, or could it be ascribed to anything? Furthermore, how could the question of responsibility be reframed in order to fit a non-anthropocentric ethical approach?

The paper focuses on the crash of a self-driving car, an example which is used to review advantages and limits of an ethical framework informed by ANT. Moreover, the article illustrates the alternative non-anthropocentric approach of Information Ethics (IE), highlighting the potentials of its narrower definition of moral agency. Ultimately, the paper shows that, despite their different ontological foundations, ANT and IE reach comparable conclusions in the moral analysis of the car crash: both these theories leave in fact the door open for the assessment of the moral agency of a self-driving system. This is possible due to a conceptual shift from the idea of *moral responsibility* to that of *moral accountability*, terms which, however, still lack a fully consistent definition.

* University of Tokyo, Graduate School of Interdisciplinary Information Studies, ITASIA D2

Key Words：Actor-network theory, Information ethics, posthumanism, moral agency, responsibility, morality of technology.