ETHICA EX MACHINA. Exploring artificial moral agency or the possibility of computable ethics

Rodrigo Sanz

Zeitschrift für Ethik und Moralphilosophie
Journal for Ethics and Moral Philosophy

ISSN 2522-0063

ZEMO
DOI 10.1007/s42048-020-00064-6

Springer

ZEMO
https://doi.org/10.1007/s42048-020-00064-6

SUBMITTED PAPER

ETHICA EX MACHINA. Exploring artificial moral agency or the possibility of computable ethics

Rodrigo Sanz

Abstract Since the automation revolution of our technological era, diverse machines or robots have gradually begun to reconfigure our lives. With this expansion, it seems that those machines are now faced with a new challenge: more autonomous decision-making involving life or death consequences. This paper explores the philosophical possibility of artificial moral agency through the following question: could a machine obtain the cognitive capacities needed to be a moral agent? In this regard, I propose to expose, under a normative-cognitive perspective, the minimum criteria through which we could recognize an artificial entity as a genuine moral entity. Although my proposal should be considered from a reasonable level of abstraction, I will critically analyze and identify how an artificial agent could integrate those cognitive features. Finally, I intend to discuss their limitations or possibilities.

Keywords Artificial Moral Agency · Normative Ethics · Machine Ethics · Computable Ethics · Moral Cognitivism

R. Sanz (✉)
Facultad de Humanidades (FHUCE), University of the Republic of Uruguay (UdelaR), Montevideo, Uruguay
E-Mail: contact@rodrigosanz.com

Published online: 06 May 2020

1 Introduction

1.1 Robots are coming

Since the automation revolution of our technological era, diverse machines or robots[1] have gradually begun to reconfigure our lives. Nowadays, no longer restricted to factories' mechanized process, they are quickly turning to new fields: social health or commercial services, stock market management, *manoeuvres* of search and rescue, entertainment, social accompaniment, transport (self-driving cars or trains), labor or military matters. With this expansion, it seems that those machines are now faced with a new challenge: more autonomous decision-making involving life or death consequences. As machines continue to increase in capacity and autonomy, human-machine interaction will be implicated by moral circumstances, inevitably. Hence, how consistently have we thought about *machine's moral behavior*?

One of the most difficult challenges in the new field of "Machine Ethics" (Allen et al. 2000, 2006; Anderson and Anderson 2011; Sullins 2006) is to confront the moral dilemmas we have been trying to answer for over 2500 years of Western philosophical thought. Thinking about this matter has been, and still is, a very difficult job, *enflaming* diverse perspectives or discrepancies about their fundamental foundations: do moral entities exist? If so, what kind of entities are we talking about or how do we get knowledge from them?

In fact, the same occurs with the exploration of a possible ethical theory for robots: what challenges or limitations are imposed since the development of an intelligence called "artificial"? Is a computational model of ethics *really* viable? Could a machine obtain the cognitive capacities needed to be a moral agent?

To answer these matters, I will expose a moral philosophical perspective focused on agency (on the cognitive characteristics of the subjects who are behind a moral action). At first, I would bring to light when we, humans, are morally considered. I will focus on a normative ethical perspective to identify when an artificial agent could be recognized as a moral entity disassociated from any biological necessity. In other words, to find the necessary and sufficient conditions in order to establish the existence of an artificial moral agent (AMA). Thereafter, I will discuss if this model could be implemented in the robotics realm.

1.2 AMA: An artificial moral agent

In the last decades, diverse opinions have been raised for the projection of an AMA. On one side, some thinkers believe robots will never have the faculty of being morals, neither today nor in the future (Bringsjord 2008). Others, conversely,

[1] "Robot" is a term commonly used today and spread mostly by science fiction literature. We are quite informal about how we use the word "robot". Its etymological origin is Czech and derives from the word "*robota*" (labor) but could also be related to the term "*rob*" (slave) of the ancient Slavic. The invention of the term is attributed to Karel Capek in his play R.U.R. from 1921. Thus, the term can be applied to a variety of artefacts: android, automaton, machine, among others. In this article, I will use indifferently "robot" or "machine" to refer to any software packaged in a hardware with a certain *degree* of artificial intelligence (AI).

consider this could be a fact even though robots are not moral agents yet (Dennett 1997). Nevertheless, skeptical pessimism aside, many people speculate their real emergence (Moor 2006; Wallach and Allen 2008; Floridi and Sanders 2004).

Technology, in some sense, has always been normative. *Artefacts* are made for different purposes and evaluated according to how well they perform their pre-assigned tasks; they are conceived taking into account their repercussions or ethical consequences. My goal here is not to evaluate an AMA by designed standards, but to examine the chances *he*[2] has to acquire some ethical and moral principles by himself. Now, determining which model could be implemented within an AMA in order to adjust to the variety of moral and social practices becomes crucial.

2 Searching for the right model: which ethical theory could be implemented?

2.1 Traditional theories

Certainly, the idea that a moral decision could be a matter of calculation is not new, in fact, it had a profound influence on the development of some traditional ethical theories[3]. Until now, engineers have handled ethics strictly as an additional set of restrictions (as other software condition). Yet, over the last decades, new studies have emerged on the subject. In 2005, a group of researchers (Anderson and Armen 2005) started a "utilitarian robotic project". Their first artificial agent, *Jeremy* (after Jeremy Bentham), deployed the "Hedonist Law of Utilitarianism" and calculates the probability of pleasure or displeasure caused on people affected by a particular action. The "right action" would be the one resulting from the best consequence amongst all possible options. But, was it really a good approach? This type of "moral mathematics" omits the strong *link* between ethics and epistemology, an oversight that brings enormous problems. Imagine the number of *infinite* potential consequences to be considered. Utilitarianism will be only *partially* computable due to the limitations of the amount of actions or the number of possible consequences that a machine can compute in order not to get stuck in an infinite process. That brings down the *basics* of this theory. Also, as pointed out by critics of utilitarianism, sacrificing a person's life for a major good is not always the best option. This type of ethical model would permanently incur in unsolvable dilemmas like the famous "trolley problem" of Philippa Foot.

As an alternative, other researchers propose to install "deontological models" (Wiegel 2006) into the machines, but again this perspective seems to forget the problematic consequences of actions followed by immovable rules in life. The model falls into paradoxes that are not easy to unravel. When a machine has specific

[2] Due to the subtle philosophical notion of "person", I will continue to use 'he' as a gender-neutral pronoun when inquiring about robot's persona.

[3] Francis Hutcheson, British philosopher of the early eighteenth century, discussed these things in "*The Manner of Computing the Morality of Actions*" (Hutcheson 1753). Hutcheson's theory is a precursor to a type of ethical theory that thrives in calculus: utilitarianism.

restrictions like Isaac Asimov's "Three Laws of Robotics" (for instance, "not injuring a human being"), given the scenario where it is only possible to save one person of a group of humans, who should be saved? A computable ethical approach based on fixed and strict rules will always be incomplete due to the struggle to anticipate all potential situations that may happen. Creating a moral robot is not merely finding the number of restrictions or right formulas to resolve conflicts, this would be a "restricted" morality and an artificial agent would result *doomed* by the limitations that someone conceived *a priori*. To reach the moral sphere, a robot must come up with *ethical principles* that could be applied to a wide range of situations and contexts instead of trying to decode them case by case: ethics as a reflection of human decision-making. But, could ethics be *resistant* to computational science?

To solve this question, I would like to present an ethical cognitivist approach that reduces many of the previous problematics when reflected inside the informatics realm. This view of normativity *depends* on the agent's position and it constitutes a path to sustain moral objectiveness, without the necessity to assume (unlike "moral realism") the ontological existence of "moral facts". One main theory is proposed by philosopher Christine Korsgaard, let's see shortly what it is about.

2.2 An ethical cognitivist theory

Korsgaard proposes a Neo-Kantian perspective based on a reflexive agency in terms of a deliberate action founded on personal identity. According to her (Korsgaard 1996), every moral agent asks himself a "normative question" in first person: why ought I to act morally? This question doesn't involve inquiring which are the specific reasons to act morally, rather, where they *come from*. And, in order to answer it, we need to identify which cognitive faculties differentiate us from other animals.

It is undeniable that numerous creatures, besides humans, have the rational abilities required for planning oriented actions to fulfil their beliefs or "first-order desires"[4], i.e., the *desire to eat*. A first-order desire is a subjective mental state which guides the agent to make an act without questioning such desire, in other words, a practical attitude addressed outside the agent. Many organisms are moved by these impulses of primary desires. However, human beings seem to benefit from a major psychological complexity. Humans may want (or not) the desires they actually have or wish other desires that will *effectively* impulse them to act. We are self-reflected beings, we can evaluate the *desirability* of our motives or desires, that is, we are capable of forming "second-order desires" (a desire that takes another desire as its object). We might, for example, "desire to eat" but also "desire to (or not) desire to eat"[5]. Humans, and only animals with high cognitive abilities, appear to have this capacity for reflective self-evaluation.

If an entity is unable to reflect about his actions – questioning them – he is not a moral nor an autonomous being; without this capacity for self-reflection he will only have desires that push him towards one direction, and later to others,

[4] The next paragraphs combine the thoughts of philosopher Harry Frankfurt through Korsgaard's view.

[5] The desires of higher or lower order do not have to be dissonant, this only shows the sinuous and complex labyrinths of the mind.

desires that exert a causal but not normative pressure on his behavior. When we act, there is *something* that is above our desires, *something* that constitutes our identity and allows us to choose which desire to endorse. Hence, according to Korsgaard (2009, 2014), only through some practical identity we can determine which desires become *our reasons to act*, and, therefore, unconditional obligations. We *are* our own normative law, and this means that the principles that guide our actions express who we *are*.

Social life implies diversity of practical identities. An individual can think of himself as a citizen of a country, a father, an academic, or even as a gangster, and what will determine how he acts is the conception he has of himself. This self-conception is not purely descriptive or contemplative but practical, it is a conception according to which we value ourselves or believe that certain actions are valuable. Humans are constituted as such when they embrace their own desires since non-compliance would imply to lose their identity, to cease to be who they are. That is why our reasons become unconditional obligations to act.

This moral perspective is not ethical relativism or subjectivism. Our reasons to act are not just private, they are rather defined in the public realm. Korsgaard's idea is that although in principle we could describe our identity from individual aspects and thereby ground how we act, in the end, the identity pilaster will always belong to the "Kingdom of Ends": we act as we would like others to recognize us (i.e. if an individual sees himself as a feminist he will never act by denying his feminist identity).

I will now briefly venture to answer if this ethical model could be computable in some type of computer system.

3 Minimum requirements for moral agency

From this cognitivist moral perspective, I will enumerate the minimum conditions under which an artificial entity would be accepted as a *genuine* moral agent. It is essential to examine the core of the theoretical model and subtract its fundamental requirements. I plan to expose them from a general view, applicable to any entity. In sum, the Korsgaardian requirements for moral agency could be:

(i) *Intelligence*
(ii) *Autonomy (as self-governance)*
(iii) *Self-reflection*
(iv) *Having at least one practical identity*

In the following sections, although my approach should be viewed from a reasonable level of abstraction, I plan to distinguish, step by step, if an artificial agent could integrate these characteristics to be recognized as a moral agent, without distorting them. Let's see if it is possible to get there.

3.1 Intelligence

Despite the word "intelligence" has different connotations, there is a common ground to all of them. The notion of intelligence is defined as a characterization in *degrees* of acquisition, that is, not all entities qualify as "intelligent". Some reach the threshold needed to be *more* or *less* intelligent, but not others (like a stone or a vacuum cleaner). Throughout history, the standard criterion turns out to be the human faculty for information processing. Human beings are the ones who evaluate and classify *intelligent* entities. But, how do we do that? What are the pragmatic principles we use to categorize them?

Inside the realm of morality, a valid answer lies in the *analysis of behavior* on the search of intelligence related to contextual practices – the ability to execute and plan behavior as *socio-spatial activities*. But still, how is it possible to characterize a mental state as a "belief" or "desire" on an entity with artificial intelligence?

Initially, I will focus my attention on a theory of the mind called "functionalism"[6]. According to it, what makes some particular mental state is not its internal constitution, but the way it works or the role it accomplishes within the system it is a part of. A mental state is a functional state, the causal role of a certain state determines what kind of mental state it is. Therefore, if a computer is reduced to processes that implement computable functions, mental states could be understood as software states.

The idea is to pay attention to how data is processed. An AMA will demonstrate intelligent behavior when he acts "intelligently" according to an observer's point of view. An intelligent machine does not have to be "identical" to an intelligent human, it should only exhibit the same type of behavior in similar circumstances. This type of behavioral criteria could facilitate the conditions demanded from robots when they *irrupt* into the field of morality[7]. Without a doubt, such task, even when exhaustively detailed, will always be in conflict with skeptics. However, it could be gradually *alleviated* if, in the near future, it weren't easy to find behavioral differences between a human being and his new existential partner with artificial intelligence.

3.2 Self-reflection (instead of self-consciousness)

In previous sections, I exposed how humans are able to endorse or question their own beliefs and desires, and that moral agency implies a reflective structure of consciousness to legislate oneself. This self-reflection is a form of rationality that has to do with the intelligent choice of ends. But, does an entity have to be self-

[6] Hilary Putnam started the debate about the nature of mental states through a functionalist lens (Putnam 1980), but, later, he also argued against it. Last Putnam's arguments try to show that functionalism cannot give an adequate account of the content of propositional attitudes (Putnam 1988). In this article, functionalism is intended to be a middle-ground approach between behaviorism and neuron-artificial-brain-specific theories.

[7] In the history of science, this type of procedure has been linked to Alan Turing, a mathematician who proposed in the 1950s an "imitation game" to determine the achievement of an AI (called the "Turing Test"). In the recent past, Allen et al. (2000) also suggested a "Moral Turing Test".

aware to be a moral agent? And if so, when do we know with certainty that an entity is *really* conscious?

The notion of "consciousness" is difficult to apprehend: there is no exact definition or a real understanding of the totality of its processes. From neurosciences to psychological or philosophical approaches, there is a vast history of diverse theories about the structure of consciousness (Gennaro 2018). A predominant common ground is to identify them with the subjective character of experience (Husserl 1962; Nagel 1974), however, to turn a subjective attribute into an objective science seems almost impossible. Epistemologically, the mental state of another "I" is never cognoscible since there are no direct means to connect with others' subjective experience, and as a result, it is only *communicable* through language or behavior. So, what are we trying to depict with the term "self-consciousness" then?

Imagine the mental life of an animal, for example, a squirrel. Presumably, squirrels have a wide variety of experiences: they feel pain or pleasure, they have a particular visual experience, they have beliefs or desires about the world around them. But even so, it is hard to think of a squirrel "aware" of itself as humans are. Is the squirrel critically endorsing its internal mental states?

According to the "Higher Order Thought" theory (Carruthers 2003; Rosenthal 2005), an experience is phenomenally conscious in virtue of *other mental state* that is about the same experience. The idea is that a conscious experience of, for example, a color, consists of a representation of such color in the visual cortex, accompanied by a high order thought in the same subject to the effect that the person is having with such color experience (Block 2009).

This type of proposal implies the ability to be aware of what is external to oneself and the faculty to access our internal states and question them. Humans are aware of their mental contents not only as an "internal experience" but as an exercise to reflect critically on them. When I are aware of our mental activity (and therefore of our behavior), it is plausible that, after thinking about it, I decide either that a behavior must change, or an action is morally reprehensible, or I made a mistake in my reasoning, or I have an unjustified belief that I should abandon.

From this perspective, to continue with the goals outlined at the beginning, human self-consciousness does not seem to be a *tacit* need for a pragmatic development of an AMA. On the other hand, "self-reflection" does, since a conscious mental state must have the *possibility* of being questioned (or endorsed) from a hierarchical level of abstraction. What I propose is that this condition sought by Korsgaard's theory (and that I am trying to mirror here within the computational field) should not be understood as "self-consciousness" but rather as a requirement for "self-reflection". If a machine manages to behave morally in a consistent manner and can respond for his actions and is able to learn from his previous behaviors, he could be a moral agent in a very real sense, *temporarily* avoiding the philosophical "problem of consciousness". Hence, how can we define, in formal terms, this self-reflection?

At a computational level, self-reflection could consist of the faculty to monitor lower level's activities, so when a behavior becomes unproductive or unsuccessful, the action is not persistently repeated; it is *identified* as useless and the agent must try something new (Hofstadter 1982). Often, a robot has a set of first-order programs that govern his most basic behaviors. Perhaps, one way of making a self-reflective

robot is to write programs whose main job is to keep track of those first-order behaviors to ensure they do not get stuck in not *perceived* circumstances. What is sought is a general "self-observation" program: a program that controls other programs but also maintains a critical eye towards its own behavior[8].

This seems to be a general rule even for us. When we are not attentive, we are vulnerable to get stuck in repetitive routines or conditioned actions. Think about daily life, how many times do we make the same pointless action before correcting it? No one is immune to that.

Clearly, a notion such as *self-consciousness* is essential for our human condition, but from a computational point of view, for an AMA, it should be seen as *self-reflection*. Isn't a computer program with "self-observation" algorithms an idea very close to it? I do not claim it is the same, in the end, only the future will reveal whether our ability to know ourselves is equivalent to a complex set of computer programs. In any case, we may think that self-observation programs will play a key role in acquiring this requisite of self-reflection, crucial for moral beings.

3.3 Autonomy: agents as intentional systems

An old common criticism in AI development is the idea that robots *only* act on how they have been *pre*coded. For sure, some programmatic coded doses are unavoidable for the construction of an operative AMA. Nevertheless, the simple follow-up of incontestable rules does not constitute a moral sovereignty but computational slavery instead. Now, the question is whether we can grant enough autonomy to artificial agents. An AMA governed only by fixed rules is not a *genuine* moral agent at all, he must build principles that guide his decision-making even when the rules come into conflict (and also to back down from them when certain circumstances require it).

Among engineers, autonomy is often addressed from its technical aspects, not from the metaphysical point of view. In many cases, they define "autonomy" as the ability to be under no direct control of any other agent or user. The more competently the robot independently achieves its objectives, the more agency is attributed to him. But autonomy thus described is not enough to obtain moral agency. Many entities such as bacteria, viruses, or even simple programs in artificial life, already exhibit such autonomy, and, of course, they should not be seen as moral agents. Freedom of action and will must arise as an emergent property in a sophisticated level of systemic organization, as it emerges in humans. Every human being is *limited* by multiple circumstances, specific to their cognitive capacity or external constraints imposed on them. People are conditioned by their past actions, their present life choices and future decisions that translate into the real capacity to implement them. Certainly, all human freedom is "restricted freedom", and so, being a moral agent does not imply "radical freedom" (as *compatibilists* have defended). Something similar happens in

[8] I am aware that this approach is related to a classic problem in computability theory: *the halting problem*. But, even if a mathematical theorem guarantees that no software will ever be a perfect observer of itself, it could simply be deduced that a perfect artificial intelligence is unattainable; something that should excite us instead of worrying us (Hofstadter 1982).

the field of the AMAs: they will have freedom within the deterministic limitations of their design.

So, how can a robot emancipate from its programmers? An interesting answer comes with the notion of "psychological autonomy" (Frankfurt 1971) since it shows that our human practices and actions can, in fact, be compatible with the belief of determinism.

Two main ideas underlie this type of model of *free will*: the psychological "division" of a person into first-and-second order attitudes, and the characterization of autonomy as a hierarchical *mesh* between second-order volitions and first-order desires. Then, where is the will exactly? According to Frankfurt (1988), a "second-order volition" is an "effective desire" that moves to act. A person can figure out any desire, but if he yearns for one to truly become his will, then he will act specifically on such first order desire rather than another. *Ergo*, personal autonomy, from this psychological point of view, resides in the hierarchical structure of desires. First, the person is constituted by identifying with some of his lower-order desires, second, he governs himself by evaluating his desires according to his own higher-order desires. Autonomy consists in the exercise of his reflexive capacity for self-identification with (some of) his desires.

In terms of the agent's own behavior, volition can be understood as the will to move by an *intentional act of the mind* because it concerns the *agent's motivational states*. Here, the will is not a *mysterious act* but a *process* that can be attributed to explain people's behavior. Yet, if *free will* could be guaranteed from this type of hierarchical theory, what would it take to compute it?

An initial approach would have to involve the distinction between different order desires. Artificial agent architectures have design methodologies. In AI research, there have been diverse perspectives for the construction of artificial intelligent agents: top-down (deliberative or cognitive), bottom-up (behavior-based or reactive), hybrid (a combination of both). Bottom-up approaches achieve their goals basically through reactions on a changing environment, which often implies nor symbolic nor representational knowledge. In fact, there is a universal acceptance that at least some amount of intelligence is best modelled with behavior-based models (Brooks 1991), but not all cognition can be systematized this way, let alone moral actions or social practices.

I do not intent to expose a full review of artificial agent architectures[9]. Yet, the combination of both methodologies increases the effectiveness of artificial systems in solving ethical problems (Bryson 2000).

A powerful modelling framework in the field of robotics engineering is BDI architecture (Rao and Georgeff 1995)[10]. Its success resides in the combination of several valuable elements: a philosophical model for deliberative reason (Bratman

[9] I am aware that there was controversy over whether behavior should be systematized using hierarchical structures (Maes 1991; Tyrrell 1993; Vereijken and Whiting 1998) or if it emerges from a more dynamic process (Van Gelder 1998).

[10] In the recent past, Honarvar and Ghasem-Aghaee (2009a, 2009b) proposed a "*Casuist BDI-Agent architecture*" which extends the power of BDI architecture.

1987), multiple possibilities of implementation, and semantics from modal logic (Georgeff and Rao 1998; Schild 2000; Haddadi and Sundermeyer 1996).

BDI systems represent a very close architecture to what I have characterized previously. It distinguishes an artificial agent with three mental states:

a. ***Beliefs*** represent the information that the agent has about the world, how he categorizes his "perception" (environment, other agents and himself). These beliefs are stored constituting a system. They can initially be programmed or acquired through experience: i.e., the fact that objects can be hard, soft or liquid.
b. ***Desires*** represent specific states of his environment that the agent would like to obtain (different from the current one). Each desire leads the agent to an *intentional* action to attain the new "state of affairs" that he seeks. In other words, desires can be understood as the agent's objectives, who must devise plans to achieve them. A simple robot could "desire" to recharge his battery, or, perhaps, he could try to make a child smile. To fulfill these desires, artificial agents must calculate on how to proceed given the conditions of the environment in his beliefs.
c. ***Intentions*** represent the desires which an agent "commit". They are the result of a desire that becomes *effective*. This means that an agent *with an intention* will begin to consider a plan to achieve an objective to which he has committed himself. These intentional desires can be simple actions, such as finding an electrical outlet, to complex strategies that require more detailed and careful planning.

Systems based on classical logical computation (oriented to perform routines) cannot recover from problems that have not been *specifically* programmed nor discover something new or take advantage of opportunities that arise unexpectedly. However, BDI architecture is a planning system that, based on models of hierarchical representation of knowledge and modal logic, allows the development of dynamic action plans in the pursuit of objectives. The system does not need to commit to fixed plans, it is capable of reconsidering them when receiving new information. Actions can be interrupted at any time by "alarm events" (i.e. he perceives an obstacle on his path). These events can reform a part of the artificial agent's beliefs, update a plan, or influence the adoption of new objectives. The implementation of plans to achieve an end constitute the concrete *intentions* of the agent, which are dependent upon his decision-making. Likewise, as humans, when we miss the bus (or train) that brings us home we still know *where* we are (beliefs) and remember *what* we want to achieve (objectives) so we can rebuild another plan. BDI systems are also capable of adjusting to changing scenarios.

The reason why I find a concrete relevance in BDI architecture is that a *desire* leads to an *intentional* action in the search for a plan to accomplish an *objective*. Here, an artificial agent could be defined as *artificial autonomous*[11] by his system of intentionality, that is, he will have practical autonomy when he acquires a symbolic model of the world (explicitly represented), hierarchical attitudes to provide him

[11] I would like to sound a philosophical alarm here. No matter how much I try to focus on the condition of self-governance, the notion of autonomy is very vast. Perhaps, a critique may emerge from the conception of autonomy regarding the capacity for reflective self-identification. I will work this aspect in subsequent sections.

with information about the environment and about himself (beliefs), *pro*attitudes that guide him to action (desires and intentions) and decision-making mechanisms using modal logic reasoning with the purpose of reaching circumstantial objectives (Molina et al. 2005; Maes 1990).

I am not suggesting that robots could achieve "radical autonomy"; in fact, humans do not need to live in a non-deterministic world to assume moral decisions. It doesn't matter if we are radically free or determined in some respects (or "programmed" for the case of robots). Being free implies psychological freedom to act. Machines will have "free will" in a similar (though not identical) way to us if their computational architecture is the "correct" one.

Now, I need to raise a self-criticism here, how can an AMA be free if he doesn't understand what "freedom" means? Furthermore, how could robots possibly *endorse* a desire to become a "normative obligation to act"? Or, even, how can an AMA attain an artificial practical identity?

3.4 Artificial personal identity

So far, part of the conditions to meet the minimum requirements of Korsgaard's theory seem theoretically viable for their implementation in the domain of robotics. If we accept the above arguments, a robot could achieve: intelligence in a rational sense of the term; self-reflection (from a functionalist perspective); and autonomy, at least from the standpoint of intentional systems and a computable hierarchical model (BDI architecture with "self-observation" programs). But, one key point about Korsgaard's moral theory is the close connection between the answers we give to the questions posed by morality and the conception we have of ourselves. Practical identities are those that provide the beliefs and desires by which we can be guided in practical life. In that case, what exactly is a personal identity in formal terms? Or, how could it be defined precisely to be computed?

According to *The Cambridge Dictionary of Philosophy* (Audi and Audi 1999), personal identity is the relationship that every entity maintains with itself. That is why a typical rhetorical safeguard is to exclaim: "*I have always been like that!*". But what is that which persists over time and which we call "identity"? According to Modern Tradition, the problem falls on what the conditions are for a person to be the same in a x time, in the following $x+1$ and in the previous x-1. Philosopher John Locke suggested that the permanence and continuity of "the same consciousness over time" (Locke 1841) would provide a good criterion for personal identity. Hence, modern derived theories are known as "psychological continuity of identity" (Parfit 1984). However, in the common rituals of social interaction, psychological continuity and physical continuity are considered together. We find it difficult to recognize a childhood friend who has changed considerably and does not resemble the figure of our memory. Although the physical or psychological traits can change, through a gradual transformation identity remains constant. If an individual grows old and loses some memories of his childhood years, he will still retain his identity, but if all his mental content is instantly emptied, what then is left of him?

This whole issue arises here to understand what relationship a notion as "episodic memory"[12] (or "Lockean consciousness") could have with practical reason. What consequences can memory have on our *reasons to act*? And what requirements are linked to it in the computable development of an artificial practical identity?

Memory is constitutive of the self-assessment processes that are intrinsically important for human beings. One of the goals of practical reason is to evaluate our existence: how our past experiences were like (positive or negative), what relationships we have with others, and so on. Memory forms an important part in the process of personal identity construction from a narrative and historical sense. It involves a self-biographical memory (Christman 2009). Without episodic memory it is not possible to get involved in an identity project of self-understanding or to highlight what could be meaningful for us in the light of the past actions' evaluation. In fact, my current condition has no meaning whatsoever outside the historical trajectory I am part of. Knowing who I am means remembering the actions and decisions with which I am involved in terms of normative obligations; it is a self-constitution over time. Therefore, identity is formed from this introspective evaluation of beliefs and desires that constitute the practical root of the individual, his system of values. I identify myself with the promises I made because I am committed to be guided by them. Without the ability to remember certain social relations or commitments, I cannot assure that my current life plans form a coherent narrative. I understand my worldly experience when I actively remember and interpret my past events as mine.

It is a historical identity[13]: a continuous process that involves a project of self-identification in relation to others. We store our social experiences similarly to how they have developed but perceived or interpreted from our perspective. These life representations are not kept in our memory simply as fixed episodes, rather, they are *impregnated* with meaning: it is an active reconstruction endowed with certain coherence, relevance and planning, enclosing the normative values that we *stamp* on them.

Surely, to program a robot to have a narrative history of "himself" or to worry about *certain things* seems viable. But, would it really be possible for a robot to question what "matters" to him as a moral reason?

Let's go back to the initial objective. From the Korsgaardian conception, a practical identity generates epistemic commitments: it represents a set of beliefs that imply morally relevant restrictions for a given identity. For that reason, individuals with different identities often disagree on their beliefs. Perhaps, a computational model for an artificial identity could be conceived as a *precoded* values system on which moral actions are justified as circumstances come up. However, as previously stated, this would not support moral principles as "normative obligations", it would only be an instrumentalization of a means to an end. Another tentative solution, as seen with BDI architecture, would be to develop the normative endorsement as hierarchical levels of "mental states" within an AMA embedded software module. Although this

[12] From a psychological-cognitive perspective, I realize that human memory cannot be simplified into a single concept. But, given the brevity of my article, and that I am linking it to personal (conscious) identity, I have decided to focus on the categorization of episodic memory.

[13] In this perspective, the influence of Paul Ricoeur (1990) should be recognized.

capability plays an important role in supporting the beliefs or as a motivational set within the agent, it fails to explain how those "reasons" are the normative source for his actions, or in what sense an AMA could really "endorse" them. At this point, the relevant philosophical question is to elucidate what could forge an AMA to "accept" the validity of his "moral reasons". Is this "adoption" the same demanded in Korsgaard's theory? Here, it seems that the Korsgaardian "endorsement" idea has another meaning.

4 Conclusions

I would like to present now two major difficulties that make me conclude that an AMA cannot currently become a genuine moral agent. Firstly, the notion of "artificial identity" carries many problems (DiGiovanna 2017). Even if an AMA could acquire all features of personal identity, this notion could be extended so largely that its original meaning would be lost. It would force us to rethink the limits of what is meant by it, because our perception of psychological and physical continuity, that adapts to our concept of identity, would vanish. In robotics, the norm is to reprogram, update and replace each of its parts. For an AMA, a radical change would be possible both physically and "psychologically". This would lead to the creation of entities that are, to a large extent, capable of rewriting and reconstituting their identities or bodies, that is, their whole being. An "artificial person" becomes too volatile: he could, by his own will, or by his designers, get rid of the fundamental characteristics that form his identity. In our social exchanges, we are used to identifing people over time, can we say the same about a being that is capable of modifying *ad libitum*? We presuppose a continuous ontic stability and we assign responsibility by identifying people's actions and consequences. If all the characteristics that constitute a person's identity are temporary or unstable, how can we trace an "artificial person" over time? This new form of identity, capable of altering his own principles or beliefs, creates enormous difficulties not only for the moral realm but also on a metaphysical level (unimaginable until today). The possibility of access and free modification of their "mental" contents entail moral dilemmas not foreseen in traditional ethical theories and alters the conditions under which moral agents operate. How can we judge them, hold them accountable, create a bond of friendship or trust them?

Even if an AMA had all the characteristics to be a person and passed all the tests philosophers have established (self-reflection, ethical cognition, autonomy, second-order volitions, etc.), he would also have a devastating feature: the ability to change instantly and without effort. If this occurs, our ethical and moral theories are conceived to blame or praise people, so which moral theory could be implemented to AMAs?

Secondly, constructing an artificial personal identity in a *narrative sense*, as a continuous process of socially mediated self-interpretation, does not seem viable yet. On the way to building artificial architectures, the research field of AI was initially divided into two paradigms: symbolic (representational knowledge, logic rules) and subsymbolic (artificial neural networks, machine learning). Nowadays, some integrated systems propose that symbolic processing functionalities could emerge from

subsymbolic structures (Schneider 2019). Yet, robotics is an early stage of maturity of the development of information theories or complex semantic processing (Floridi and Sanders 2004; Floridi 2004; Taddeo and Floridi 2007). The importance of this point can be glimpsed by grasping the impossibility to *unlink* the moral domain from epistemology. Undeniably, making judgments requires information that describes the *factual* states of any normative evaluation. So, how could an AMA possibly process the immensity of data that comes from the outside world to act and reflect morally?

Bernd Carsten Stahl thinks that when "data acquires context-dependent meaning and relevance, it becomes information" (Stahl 2004). That is, I have "information" when I have "data with meaning":

> We are surrounded by a potential infinity of data, by reality as it is, which we have to endow with meaning in order to be able to make use of it. (...) Agents, whether human or not, must take their perceptions of reality in order to identify the data that is relevant (...) [and] endow the data with further meaning necessary to act on it. (Stahl 2004: 73)

Information allows us to perceive data in such way that it acquires meaning. Information constitutes meaning. A stone only becomes a stone when it is perceived as such and only then it can be used. Even if you don't really know the "name" of it, you need to have some information to employ it as a weapon or as a tool (to construct a meaning). Hence, in order to actively participate in normative practices, an AMA must have a complete vision of the data set that needs to be organized from an active engagement, situating him in a timeline that connects him with the environment in a meaningful way. To make moral judgments, he needs to prioritize between the data that arrives to his artificial senses. This includes the ability to decide which data is relevant for processing and which is not (most cognitive ethical theories demand this regardless of their theoretical origin). Instead, nowadays, robots are not *skilled* for assigning "meaning" to the data they process:

> There are no algorithms to decide a priori which data is relevant and which is not. Human beings are able to make that decision because they are in the situation and they create the relevant reality through interaction. Computers, on the other side, do not know which part of the data they process is relevant for the ethical evaluation because they miss their meaning. (Stahl 2004: 78)

Robots, in their present condition, are not *information processors* in a relevant cognitive sense, but rather extraordinary *data processors*. However, over the last decades, research on integrated neural-symbolic systems has made a significant progress in the attempt to overcome these limitations. Neuroscientists, philosophers and psychologists are providing new theories and models that represent a rich source of inspiration for the field of AI. In particular, cognition theories that enable new types of algorithms and architectures (Laird et al. 2017; Hassabis et al. 2017). While this exploration is still in its early stages, there are some good examples with promising guidelines for future research (Choi and Langley 2018; Borst and Anderson 2015; Trafton et al. 2013; Laird 2012).

Clearly, the challenge of ethical computing is related to a more general research of an epistemology of artificial systems and learning methods. To be part of the

moral sphere, artificial agents should be capable of learning by themselves from diverse social interactions in order to improve their moral behaviors and respond judiciously to situations they have never encountered before. The problems of symbolic paradigms will never perish without a situated knowledge. Likewise, if robots are deprived of internal symbolic representations, they will hardly acquire a cognitive and contextual meaning of their environment.

To understand information (moral or not) in a "meaningful" way, robots will have to be part of human life, actively participate in moral discourses and acts, and contribute within moral communities. Meaning emerges from a social construction that results from diverse pragmatic interactions. As Korsgaard assert, any reason to act is a *public* reason, it comes from a socially shared language and interactive experiences. Being a moral being is linked with the ability to act prospectively: to deal with what *might be*, not just with *what is*. A robot will never act as a *genuine* moral agent until he achieves a real connection with our social practices, since we become moral agents through socialization, enculturation and learning.

References

Allen, Colin, Gary Varner, and Jason Zinser. 2000. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence* 12.3: 251-261.

Allen, Colin, Wendell Wallach, and Iva Smit. 2006. Why machine ethics? *IEEE Intelligent Systems* 21.4: 12-17.

Anderson, Michael, and Susan Leigh Anderson, eds. 2011. *Machine ethics*. Cambridge University Press.

Anderson, Michael, S. Anderson, and Chris Armen. 2005. Towards machine ethics: Implementing two action-based ethical theories. *Proceedings of the AAAI 2005 Fall Symposium on Machine Ethics*.

Audi, Robert and Paul Audi (Eds). 1999. *The Cambridge dictionary of philosophy*. Vol. 584. Cambridge: Cambridge university press.

Block, Ned. 2009. Comparing the major theories of consciousness. *The cognitive neurosciences*. 1111–1122. Cambridge, MA: MIT Press.

Borst, Jelmer P. and John R. Anderson. 2015. Using the ACT-R Cognitive Architecture in combination with fMRI data. *An introduction to model-based cognitive neuroscience*. Springer, New York, NY, 2015. 339-352.

Bratman, Michael. 1987. *Intention, plans, and practical reason*. Vol. 10. Cambridge, MA: Harvard University Press.

Bringsjord, Selmer. 2008. Ethical robots: the future can heed us. *Ai & Society* 22.4: 539-550.

Brooks, R. A. 1991. Intelligence without representation. *Artificial intelligence*, 47(1-3), 139-159.

Bryson, Joanna. 2000. Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental & Theoretical Artificial Intelligence* 12.2. 165-189.

Carruthers, Peter. 2003. *Phenomenal consciousness: A naturalistic theory*. Cambridge University Press.

Choi, Dongkyu, and Pat Langley. 2018. Evolution of the ICARUS cognitive architecture. *Cognitive Systems Research*, 48: 25-38.

Christman, John. 2009. The politics of persons: Individual autonomy and socio-historical selves. Cambridge University Press.

Dennett, Daniel C. 1997. When Hal Kills, Who's to Blame? Computer Ethics. *Hal's Legacy: 2001's Computer as Dream and Reality*, 351-365. Cambridge, MA: MIT Press.

DiGiovanna, James. 2017. Artificial Identity. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*.

Floridi, Luciano. 2004. Open problems in the philosophy of information. *Metaphilosophy*, 35, 554–582.

Floridi, Luciano and Jeff W. Sanders. 2004. On the morality of artificial agents. *Minds and machines* 14.3: 349-379.

Frankfurt, Harry. 1971. Free Will and the Concept of a Person. *Journal of Philosophy* 68.1: 5-20.

Frankfurt, Harry G. 1988. *The importance of what we care about: Philosophical essays*. Cambridge University Press.

Van Gelder, Tim. 1998. The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences* 21(5), 615-628.
Gennaro, R. J. (Ed.). 2018. *The Routledge Handbook of Consciousness*. New York: Routledge.
Georgeff, M. and A. Rao. 1998. Rational software agents: from theory to practice. *Agent technology*. Springer, Berlin, Heidelberg. 139-160.
Haddadi, Afsaneh, and Kurt Sundermeyer. 1996. Belief-desire-intention agent architectures. *Foundations of distributed artificial intelligence*. 169-185.
Hassabis, Demis, D. Kumaran, C. Summerfeld, and M. Botvinick. 2017. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2), 245–258.
Hofstadter, D. R. 1982. Can creativity be mechanized. *Scientific American* 247.3: 18-34.
Honarvar, A. R. and N. Ghasem-Aghaee. 2009a. An artificial neural network approach for creating an ethical artificial agent. *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation-(CIRA).* 290-295. New Jersey: IEEE.
Honarvar, A. R. and N. Ghasem-Aghaee. 2009b. Casuist BDI-agent: a new extended BDI architecture with the capability of ethical reasoning. *International conference on artificial intelligence and computational intelligence*. 86-95. Springer, Berlin, Heidelberg.
Husserl, Edmund. 1962. *Ideas: General introduction to pure phenomenology.* 1913. Trans. WR Boyce Gibson. New York: Collier.
Korsgaard, Christine M. 1996. *The sources of normativity*. Cambridge University Press.
Korsgaard, Christine M. 2009. Self-constitution: Agency, identity, and integrity. Oxford University Press: USA.
Korsgaard, Christine M. 2014. *The constitution of agency: Essays on practical reason and moral psychology*. Oxford University Press: USA.
Laird, J. E. 2012. The Soar cognitive architecture. Cambridge, MA: MIT Press.
Laird, J. E., C. Lebiere and P. S. Rosenbloom. 2017. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38(4), 13–26.
Locke, John. 1841. An essay concerning human understanding.
Maes, Pattie. 1990. Situated agents can have goals. *Robotics and autonomous systems* 6.1-2: 49-70.
Maes, Pattie. 1991. A bottom-up mechanism for behavior selection in artificial creature. *Proceedings of First International Conference on Simulation of Adaptive Behavior (SAB'90).* Cambridge: The MIT Press
Molina, José Manuel, Juan Manuel Corchado Rodríguez, and Juan Pavón Mestras. 2005. Modelos y arquitecturas de agente. *Agentes software y sistemas multiagente: conceptos, arquitecturas y aplicaciones*. New Jersey: Prentice Hall.
Moor, James H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems* 21.4: 18-21.
Nagel, Thomas. 1974. What is it like to be a bat? *The philosophical review*, 83(4), 435-450.
Parfit, Derek. 1984. *Reasons and persons*. New York: Oxford University Press.
Putnam, Hilary. 1980. The nature of mental states. *Readings in philosophy of psychology* 1: 223-231
Putnam, Hilary. 1988. *Representation and reality*. Cambridge, MA: MIT Press.
Rao, Anand S. and Michael P. Georgeff. 1995. BDI agents: from theory to practice. *ICMAS*. Vol. 95.
Ricoeur, Paul. 1990. *Soi-même comme un autre*. Paris: Le Seuil.
Rosenthal, David. 2005. *Consciousness and mind*. UK: Oxford University Press.
Schild, Klaus. 2000. On the Relationship Between BDI Logics and Standard Logics of Concurency. *Autonomous Agents and Multi-Agent Systems* 3, 259–283 https://doi.org/10.1023/A:101007602770.
Schneider, Howard. 2019. Emergence of belief systems and the future of artificial intelligence. *Biologically Inspired Cognitive Architectures Meeting* (pp. 485-494). Springer, Cham.
Stahl, Bernd C. 2004. Information, ethics, and computers: The problem of autonomous moral agents. *Minds and Machines*, *14*(1), 67-83.
Sullins, John. P. 2006. When is a robot a moral agent. *Machine ethics*, 151-160.
Taddeo, M. and Floridi, L. 2007. A praxical solution of the symbol grounding problem. *Minds and Machines*, 17(4), 369-389.
Trafton, J. G., L. M. Hiatt, A. M. Harrison, F. P. Tamborello II, S. S. Khemlani and A. C. Schultz (2013). Act-r/e: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction* 2(1), 30-55.
Tyrrell, Toby. 1993. The Use of Hierarchies for Action Selection. *Adaptive Behavior* 1(4), 387–420.
Vereijken, Beatrix and H. Whiting. 1998. Hoist by their own petard: The constraints of hierarchical models. *Behavioral and Brain Sciences* 21(5), 705-705.

Wallach, Wendell and C. Allen. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.
Wiegel, Vincent. 2006. Building blocks for artificial moral agents. *Proc. Artificial Life X*.