Human Beings and Robots: A Matter of Teleology?

Andrea Lavazza †
lavazza67@gmail.com

ABSTRACT

In this paper, I use the comparison between human beings and intelligent machines to shed light on the concept of teleology. What characterizes human beings and distinguishes them from a robot capable of achieving complex objectives? In the first place, by stipulating that what characterizes human beings are mental states, I consider the mark of the mental. A smart robot probably has no consciousness but we might have reason for doubt while interacting with it. And a smart robot shows intentionality. I focus on the type of naturalized intentionality that is at stake here. Then I go back to the traditional idea of teleology, and to the scientific criticism of it, through the question of the kind of purposes that artificial intelligence (AI) may set itself. Husserl′s basic idea of teleology therefore serves to have an authoritative term of comparison and to introduce the intuitive difference between human beings and intelligent machines based on the *homo pictor* thought experiment proposed by Jonas. My conclusion is that a specific finalism, understood in a non-criterial sense, is what qualifies the human being and differentiates the latter (for now) from smart robots.

1. Introduction

The European Parliament has recently expressed the hope that smart robots will be given some sort of legal status. The latter are defined as ″machines″ that have the following characteristics:

- the acquisition of autonomy through sensors and/or by exchanging data with their environment (inter-connectivity) and the trading and analyzing of those data
- self-learning from experience and by interaction (optional criterion)
- at least a minor physical support
- the adaptation of their behavior and actions to the environment
- absence of life in the biological sense (EU, 2017)

† Centro Universitario Internazionale, Arezzo, Italy.

The aim is to introduce a legal framework for such machines, as they now have a high degree of direct interaction with human beings and often partly or totally replace human operators in the implementation of particularly delicate functions and tasks in terms of security or privacy. In the Resolution is stated that the goal is to create "a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions" (EU, 2017).

The attempt to proceed to a partial assimilation of intelligent machines to human beings in terms of legal status and assignment of responsibility has caused perplexity and protests (Nevejans *et al.*, 2018). In this political diatribe, what is philosophically interesting is the possibility of identifying more precise criteria on the basis of which to differentiate or liken intelligent machines to human beings.

In fact, the criteria introduced by the European Union seem rather unsatisfactory and non-comprehensive. Now, a univocal and commonly accepted indication in this sense could be the Turing test: a machine that passed that test would in fact be by definition indistinguishable in its communicative interaction from an average human being. However, no machine has passed the test convincingly so far. It may therefore be useful to analyze what makes the human being intuitively unique as opposed to smart machines - or, alternatively, analyze why there is some similarity between human beings and intelligent machines - thereby treating the issue in a philosophically more sophisticated way.

2. Consciousness and /or Intentionality

One may think that what qualifies the human being as such compared to a robot that, say, can beat her at chess are her mental states (I will not say that the human being has a ″mind″ because the latter seems belong to an ontology that is very controversial today). Some might argue that the computational states of the most sophisticated machines are not cognitively different from human mental states. This requires an analysis of what qualifies the mental states of an average human being, including their teleological structure (cf. Lavazza, 2015).

There is sufficient agreement on the fact that mental activities or states amount to feeling (sensations), perceiving, thinking and being conscious. For Descartes, thought is everything that we are aware is happening within us (Descartes, 1641). The salient aspect of our mind would therefore be consciousness: William James (1890) believed that the presence of a consciousness of some kind was the first and most concrete fact that anyone could attribute to their own inner life. A state (or an event) would therefore be mental if it is conscious, but today this idea is no longer sustainable, given that we now know that much of mental activity falls within the sphere of the so-called cognitive unconscious (Bargh, 2017). Many mental processes (a series of states and events connected to a specific activity: for example, recalling a memory) are fragmented and carried out on a sub-personal level, of which we are not aware at all.

Mental phenomena - an umbrella term for states, events and processes - are therefore complex and heterogeneous, as they can be identified differently on the basis of different criteria. A relevant bipartition is that between qualitative-phenomenal and intentional aspects. The former refer to a private, subjective conception of mental phenomena; the latter to a public conception that focuses on intentionality, mental representations and propositional content. According to Kim (2011), qualitative mental phenomena include:

1) *Sensory sensations or qualities:* pain, tickling, seeing a yellow stain, tasting chocolate ice cream, feeling nauseous. These are states that have a phenomenal character, and are marked by the subjective experience they arouse.

2) *Emotions* (for example: joy, anger, fear, disgust) and *feelings* (for example: disappointment, remorse, pride, shame). Some may have a propositional content: you may be embarrassed for having broken something in somebody′s house.

In this first qualitative sense, one can say that an organism has conscious mental states if "there is something that it is like to *be* that organism - something it is like *for* the organism" (Nagel, 1974). None of this would happen, presumably, for a hair, a chair or a leaf. The fundamental fact is that we have non-inferential knowledge of conscious mental states: that is, when we are conscious, we know it, and we know this directly, without needing to rely on some other (prior or different) source of knowledge.

Intentional phenomena instead include:

1) *Mental representations*: internal states that, for example, ″stand for″ the properties of the objects we experience, express a certain state of affairs, constitute contents to which thoughts relate. Representations can be images, like those aroused by vision, but generally they do not have to be understood in a figurative sense. Having a belief about the seven hills of Rome means saying that one thinks of Rome as having seven hills, even though ″representation″ is a general term (a thermometer also ″represents″ the temperature with its mercury levels).

2) *Propositional attitudes*: (intentional) mental states, such as desires and beliefs, attributed to a person that, having or exemplifying them, expresses an attitude towards a proposition (that is, a content). In other words, a person may "*hope* that her boyfriend wants to get married", "*believe* that his boss is not good enough for his role" and "*doubt* that her favorite team will win the match". Propositional attitudes are a fundamental component of intentional psychology and it is generally believed, with some exceptions, that they do not show qualitative aspects. Intentional mental phenomena also include:

3) *Volitions*: to have an intention, to want, to decide. They have a propositional content and are closely related to actions. According to many philosophers, all actions are preceded by an act of volition because actions in the intentional sense are not simple bodily movements (but this is a controversial topic).

4) *Character traits or personality traits, habits, propensities, intellectual abilities*: they are usually considered mental in an indirect or derived sense, that is, as dispositions or tendencies to form wishes of a certain type and to act accordingly.

5) *Rationality*: understood as the possibility of classifying behavior, described in terms of intentional mental states and explained by referring to causes, reasons and criteria of coherence.

3. The Mark of the Mental

One may therefore wonder what characterizes human mental properties: in other words, what is the *mark of the mental*? This is relevant because the mental is the level where one can equate or differentiate between human beings and intelligent machines. Despite the many proposals available, there is no agreement on a single necessary and sufficient condition that defines it. In any case, the main used criteria are epistemological: the difference between phenomena lies in the way we know about properties. However, there must also be ontological criteria, referring to what exists in the world and its nature. To sum up, the mental in a qualitative and subjective (private) sense - as opposed to the physical sense - is characterized by:

1) *immediacy*: direct introspective knowledge of our internal states not based on external evidence or inferences (that is, reasoning based on beliefs or factual

data, unfolding from premises to conclusions). For example, it is quite inappropriate to ask someone: ″How do you know you have a toothache?″;

2) *privileged access and first-person perspective:* mental states are given in a unique way to those who experience them, they appear to us ″in the first person″, with a character of ″privacy″. As Frege (1918) pointed out: "everyone is presented to himself in a particular and primitive way, in which he is presented to no-one else". The first-person perspective, that is, one′s own point of view on things, is that which differentiates subjectivity from third-person scientific objectivity, for which there are no special points of view in the quantitative, replicable and referable description of reality. A toothache is presented to the person who has it - nobody else can experience ″that″ toothache. The asymmetry between first-person and third-person knowledge is evident when it comes to pain, but it can also exist with perceptions, given that everyone can react differently to seeing a given color. Perspective is a condition for mental states, but not a mental state in itself;

3) *immunity from error and transparency with regard to the self-ascription of mental states*: a mental state that we feel is ours can only be ours - unlike, for example, the content of a memory. Mental states are ″incorrigible″. Mental events (starting with pain) are such that one cannot be mistaken about them. The transparency of mental events implies that if an event occurs, the subject is aware of it. Infallibility and transparency of the mental are Cartesian characters, which however seem to present several exceptions. There are in fact many unconscious mental states, which we deny having and yet strongly influence our behavior (starting from prejudices, as we well know after Freud and the discovery of cognitive unconscious) and there are behaviors for which we give rational explanations when instead they are produced by automatisms or cognitive distortions;

4) *phenomenology*: it is a question of what it feels like to be oneself or to experience certain sensations. This is distinct from privileged access, which is a purely epistemological criterion. It means that we react to something in a peculiar way, that we have a special interaction with the object; that the object has qualities that only we know how to grasp; that we resonate in peculiar ways without there being a fixed correspondence between object and reaction. In the field of transcendental phenomenology as Husserl understands it, finalism is an important element, because the life of consciousness has an all-pervasive teleological structure: the different layers of constitutive syntheses can be described as pointing to an ultimate goal;

5) *unity*: the stream of consciousness and the focus of attention are convergent and indivisible; only serious pathological situations (such as schizophrenia) prevent a clear inner unitary sensation. However, many philosophical objections to the unity of the mind have been raised. The first was David Hume′s theory, according to which introspection only reveals bundles of sensations and perceptions unrelated to one another.

On the other hand, in the definition of intentionality adopted by the philosophy of mind, as explained and updated by Brentano (1884), it is stated that the true peculiar characteristic of all mental states is intentional inexistence. This thesis is today embraced by various scholars, including for example Crane (2001): according to him, the directionality of the mind towards its objects is the distinctive and exclusive sign of mental phenomena. Mental states (all of them, according to Brentano; not all of them, according to many philosophers) always have a content: they refer to, have as their object, or pertain to things, or states of affairs in the world, other than themselves. In other words, they have a mental element with representational properties, a content that does not necessarily exist - think e.g. of a hippogriff - but is such as to render these states semantically evaluable as true or false. It can therefore be said that intentional objects (concrete, indeterminate or non-existent) do not have a nature of their own, but can be defined real in the sense that their intentional object has a reference.

One can distinguish between a referential intentionality (which concerns the orientation of our thoughts: when we think of ″the Colosseum″, we refer to the Colosseum) and a content intentionality (which concerns the class of mental states - i.e. propositional attitudes - that have a content or meaning expressed by a proposition, such as, for example, ″I believe it will rain tomorrow″). It is because they have a content that mental states represent certain states of the world, that is, they have the capacity to represent entities external to them. And if intentional mental states have objects, they present them in a specific way, according to a personal perspective or with an aspectual shape. So, you can think of London imagining it covered in snow, or filled with sunshine in summer, or according to the memory of a postcard you once received. This means that there is no pure reference of mental states. Every mental access to objects is marked by the perspective from which it is considered. In this way, the intentional object is defined in terms of directionality, while the intentional content is defined in terms of an aspectual shape (Crane, 2001). Finally, the intentional mode is the relationship between the thinking subject and the contents of intentional states: one can hope that it rains, or one may believe it will rain.

Another distinction is that between *intrinsic or primary intentionality* and that which is ″transferred″ in produced or interpreted symbols (writing or a computer program, which also have a meaning and a reference). This one constitutes a *derivative intentionality*, because it depends on the interpreter. John Searle has also spoken of an ″as-if intentionality″, when something exhibits an intentionality that it does not have. The mental experiment of the Chinese room (Searle, 1980) is linked precisely to this. Let′s say that inside a closed room there is an English-speaking subject who speaks no Chinese. He is provided with an English manual - a computer program - based on which he can answer correctly, with Chinese symbols (which he finds in appropriate boxes), to other symbols that, unbeknownst to him, express questions asked in Chinese. The subject′s answers will appear perfectly sensible to an external observer, to the point of claiming that the person in the room speaks Chinese; in reality, the subject is only using symbols according to syntactic rules, but does not understand their meaning. The argument aims to show that a computer is limited to syntax (that is, rules of composition) but has no access to semantic content (meaning), which is an exclusive ability of the human mind. However, says Searle, even computation and syntax relate exclusively to the observer: there are no intrinsic or original computations in nature.[1]

Ultimately, it can be said that mental states must be at least conscious or intentional: there are indeed dispositional beliefs (what we know but we are not thinking about, like the concepts stored in memory) and states like anxiety, which are conscious but not intentional. If we also consider that many unconscious states could also become conscious, it seems that the mark of the mental is phenomenal consciousness (as claimed, among others, by Galen Strawson, 1994). If instead one adopts the view that the whole consciousness is representational (which is controversial), the unitary mark of the mental seems to be intentionality (as proposed by Crane, 2001).

[1] Searle′s argument has given rise to a broad debate and, like many other well-known thought experiments, is very controversial; the most widespread objection calls into question the system as a whole (room, subject, symbols, manual): thus conceived, the system would understand the meaning of the symbols. According to the author, however, one can place all these elements in the person′s head, memorizing them mechanically, and the latter will still not speak Chinese. Today very few scholars worry about Searle′s argument and assume that computers will be increasingly smarter.

4. Artificial Intelligence and Intentionality

One of the many concepts for which it is difficult to find a shared definition is that of intelligence. A proposal that has various supporters equates intelligence with the ability to achieve complex goals. Artificial intelligence is a non-biological intelligence that is currently passing from a ″restricted″ phase, in which it has the ability to reach a limited set of goals - from playing chess to driving a vehicle - to a general phase, in which it will be able to achieve any goal, including learning. Software that vocally respond to our commands and lead us gently to our destination, or robots capable of looking after an elderly person, giving the impression of caring of them in an almost human way, are the candidates for the new legal status currently under discussion in the European Union and in the philosophical arena. As said earlier, giving intelligent machines a similar status to that of the human being is an interesting philosophical issue. In fact, this comparison cannot solely rest on a general executive functionality. In reality, artificial intelligence has already overcome the executive efficiency of biological intelligence in many respects. The latter remains superior when it comes to interactions that specifically concern the human world, thus demonstrating a peculiarity.

The mental component of the human being, as we have seen, is characterized by consciousness and / or intentionality. As for the ability to have subjective experiences, there seems to be broad consensus on the fact that smart machines today do *not* have primary sensibility, although they can be instructed to verbally express and visually mimic feelings through the physical medium they are endowed with. An artificial intelligence cannot feel anything like what a human feels when seeing a red wall, tasting chocolate ice cream, or being caressed - and not just for the lack of an adequate material connection with the world. Nor do computers seem to have any basic awareness of their own existence and functioning, understood as a second-level computation as opposed to first-level functionalities. But what practically matters in our relationship with new generation machines is above all their ability for competent interaction: the intelligence shown by the software.

However, according to an eliminative perspective *à la* Dennett (2017) or to a logical-behavioral view *à la* Ryle (1947), one may think that a very intelligent robot could still deceive us about its (true or presumed) consciousness. This would not be a deception wanted by the robot, but an intrinsic difficulty in detecting pure conscious states, separated from more general cognitive states. In fact, based on the characteristics listed in section 3, the mental in a qualitative and subjective sense is private and difficult to grasp for an external observer.

First of all, immediacy is a characteristic that we could not challenge in a robot any more than in a human being saying, for example, ″I feel (am) hot″. The first-person perspective is often thought not to be a primary and distinctive fact, and it would be hard to tell if a smart robot has something like that, regardless of what it may say (in fact Wittgenstein could say that even an advanced robot would be able to learn from human beings to react to certain situations in typical ways and then learn to define certain states in a shared way). Nor can a robot be mistaken about its internal states, which could be characterized by some output unification, if necessary.

Phenomenology (what is it like to), however, is a different story: as mentioned, it can be assumed that machines do not have it, but being something elusive, an advanced artificial intelligence could possibly give us the impression of feeling something. Given this elusive character of the component of consciousness, the fact remains, as already noted, that in its interactions with human beings the behavior of a smart robot seems oriented by an autonomous intentionality. But can artificial intelligence be endowed with an intentionality comparable to that of human beings?

4.1. Naturalized Intentionality

Primary higher-order intentionality has always been considered a prerogative of the human being. So far, no theory has succeeded in giving an adequate account of it, overcoming potential objections. As is known, approaches have been recently proposed that use teleological categories to try to explain the problem of intentionality. *Biological theories* (e.g. biosemantics, see Millikan, 1984) claim that intentionality is a purely natural function. Just as the lungs, formed under the pressure of evolution, pump oxygen into the organism, so the function of *wanting to feed* causes an organism to seek food in order to survive and reproduce. Regarding that mental state, having food as one′s intentional content plays the role of inducing the body to feed itself. Meaning, in this perspective, is equivalent to the function performed, in the way in which it was selected by evolution according to environmental pressures. The biological function of a phenotypic trait has a purpose acquired along the evolutionary path; in the same way, beliefs would have been selected to bring information about the environment (teleonomic conception). A mental state that tends to be caused by a lion (e.g. fear) is useful for survival and is therefore replicated and transmitted.

According to the proponents of the biological theory of intentionality, intentional content is safe from the problem of misrepresentation, given that if it derives from the function it must perform, it will take place even if the cause of the content is different from the typical one (the propensity to avoid hot objects has this meaning because it serves the function of ensuring that the body does not injure itself, and its meaning remains unchanged even when a specific instantiation of that mental state is provoked by an object painted red and only apparently red-hot). However, this family of theories faces strong difficulties when it comes to explaining complex and abstract intentional contents or concepts of recent formation, such as "Raphael is a better portraitist than Pollock" or "ions" and "CEOs". The answer is usually that such mental states can be derived through a secondary functionality, thanks to their relationship with basic mental states, but this general statement is hard to be clearly detailed to account for how these processes occur.

Teleological functionalism only ascribes mental states to systems that are organized in a teleological way, i.e. literally oriented to a purpose or implementing a goal-directed behavior. In this sense, a paradigmatic case is intentional human action, in which the behavior of the subject derives from a decision and a choice. However, teleological functionalism is necessarily more liberal and also accepts behaviors or processes that are oriented to an end, although they do not imply a decision and a choice (cf. Jaworski, 2011: 53). For example, plants turn towards light sources to obtain the energy necessary for photosynthesis. It can be said that the phenomenon of phototropism is oriented to this purpose, but not that the plant decides to obtain energy and, considering the various means available to do this, chooses one over the others. What happens is that the different components of a complex system seem to cooperate in the achievement of a goal.

The difference between artificial systems and natural systems like an organism seems to lie in the way in which teleology is determined or emerges. The teleology of an artifact is, in general, programmed by its designer, while in natural systems teleology is believed to be due to natural selection which, in turn, can be potentially be seen as oriented to a higher-order goal than individual organisms. In artifacts, it is the designer who conceives a purpose for the artificial system and makes it or programs it in a way that fits that purpose. In the case of natural selection, there is a process of self-organization that induces the system to pursue goals which in turn may be instrumental towards more general purposes or be an end in themselves. According to teleological functionalism, mental states only exist in systems that are organized in a teleological way. The argument is designed to prevent functionalism from leading to paradoxes, such as the claim that even a swarm of bees moving in a coordinated (and apparently finalized) manner is a system endowed with mental states. However, teleological functionalism may not be enough to refute examples such as Block′s Chinese Nation argument.

4.2. Machine Learning

Artificial intelligence, whether or not it is considered endowed with primary intentionality, is capable of learning concepts, albeit still rudimentary ones, thanks to cognitive architectures that mimic the neural networks of the human brain. Human beings do many things easily, generally thanks to unconscious and subpersonal processes, of which we are not aware in their computational junctions. So, we are not able to come up with (algorithmic) instructions on how to recognize a face after twenty years since the last time we saw it, how to identify the style of a famous painter in the framework of another artist, how to understand whether someone is lying to us and so on. Smart machines, thanks to refined computational processes implemented in artificial neural networks, can instead recognize faces, replicate the style of an artist and discriminate between subjects who are lying and subjects who are telling the truth.

Non-supervised deep learning, used when the environment is extremely complex and changeable, employs algorithms that follow no guideline: they are free to roam through large data fields looking for some ″order″, consisting of clusters in which the algorithm distributes the elements it treats based on characteristics it recognizes as similar or assimilable (cf. Greenfield, 2017). It can, for example, create clusters of what we call ″musical genres″ by analyzing many different pieces, even without any preliminary indication. The greater the amount of available data, the greater the possibility for the algorithm to identify characteristics and elements of classification and to find regularities in its environment. It therefore constructs profiles of things that appear to be persistent and salient within a given time unit. And there is a convergence towards what is salient and relevant also for the actors the algorithm must interact with, especially human beings.

5. Smart Machines and Their Purposes

In the reflection on artificial intelligence, the topic of the ends or purposes of artificial intelligence is particularly important. This leads in an interesting way to a consideration of purposes from a renewed philosophical and also metaphysical perspective, which in many other respects appears neglected today. The discussion usually focuses on the purposes that machines may be able to set themselves in the future, goals that could diverge from human ones, putting creatures against their creators, so to speak, with the former being much stronger than the latter at that point. Here, however, I will focus on the quality of these goals from a more philosophical standpoint.

5.1. The Parable of Teleology

Following the influential view of philosophical biology defended by Hans Jonas (1966, chap. 2), we can reconstruct a macro-path that starts from an animistically undivided reality, passes through a reality divided into two kingdoms, and then reunites them under the sign of naturalization. The starting point is given by the (probably innate) tendency to read reality in a teleological way and to attribute a ″meaning″ to natural phenomena and processes in line with to human affairs, an inclination that already shows in preschoolers and remains strong even in educated adults (Kelemen, 1999; Kelemen and Rosset, 2009). Seeing a purpose in nature or trying to find a purpose in what happens leads to dissatisfaction with explanations that by definition avoid asking ″why″ and rather ask ″how″. In other words, what is sought is an explanation that goes beyond a simple mechanistic-causal account of phenomena. Common sense is more at ease in an anthropomorphic framework that starts from direct observations and then infers unobservable layers that place the world within the plot of a coherent explanation, which is not solely entrusted to chance or ″impersonal″ laws (cf. Nagel, 2012).

The ordinary perspective, in fact, tends to reject physical, biological and evolutionary determinism, preferring a voluntaristic vision of the human world, in which people are not at the mercy of events and circumstances, as this latter view somewhat implies a degradation of their dignity (often there is also a dimension of fatalism, but widespread intuitions and beliefs are not always consistent). Not even the acquisition of a scientifically informed perspective on physical reality and the functioning of the human being seems to entirely overshadow the ordinary and spontaneous teleological perspective, based on the ability to act in view of a goal. A recent study (Kelemen *et al.*, 2013) has shown that not only ordinary people, but also academics of humanistic subjects and even professional scientists agree on clearly unfounded teleological statements, such as ″The sun radiates heat because warmth nurtures life″, when called to respond in a very short period of time, which does not allow to resort to reflective evaluation.

The first passage identified by Jonas is to remove final causes from nature, in accordance with the presuppositions of the scientific method and its philosophical premises. The scientific method places the controlled observation of phenomena at the center of cognitive research through repeatable experiments, which fragment the phenomena into clearly identifiable portions. With the Cartesian ontological and epistemological turn, anthropomorphism remained confined to the world and sciences of life, and to the human being in particular - the only area where finalistic explanations were still admitted. Even in classic machines one could identify a teleology imposed by man, namely the final cause that the designer inscribed in his project, but according to Jonas the artifact′s functioning is only seen in the efficient causes that translate its invisible final causality.

The tendency to grasp a certain finalism in the world, as seen, is strongly rooted in our basic conceptual categories, and since Francis Bacon and his *idola tribus* it has been considered an innate prejudice. Bacon himself believed that final causes were something that is more about the nature of man than the universe. And this view was crystallized by Descartes into an ontological principle - a metaphysical assumption more than a scientific fact: reality is divided into *res extensa* and *res cogitans*. The former became the realm of the application of mathematical and mechanistic analysis, where finalism is substantially banned from, due to the prevalence of an objectification that rejects subjective and vitalistic aspects. The latter were attributed exclusively to the dimension of man and his constitutive freedom. Anthropomorphism represented the negation of scientificity and even efficient causes were questioned by the Humean conception.

The original finalism was thus confined to the human sphere - which, however, included the organic dimension of the human body. With the progressive affirmation of the materialist paradigm and then of the Darwinian one, the latter could no longer be conceptualized as distinct from the rest of the non-organic world, as a result of which the dualistic perspective was called into question. The *res cogitans* was progressively absorbed into the physical domain. Thus the unique and peculiar characters of first-person subjectivity that characterized the Cartesian mind and gave it a specific status, one in which teleological explanations still had citizenship, were lost. The evolutionary mechanisms of random mutation and selection modeled by the environment, in fact, removed every possibility of a final cause from the dynamics of life. Hence a return to a unity of the real under the naturalistic perspective, which has no space for teleology.

At this point, says Jonas, it is at best acceptable to hypothesize a general theistic or immanent end (the anthropic principle) which, however, is realized through local afinalistic causes. This may help give meaning to reality and its becoming, but it cannot be used as an explanation within the scientific discourse and the scientifically informed philosophical discourse.

5.2. Back to a Naturalized Teleology

In a naturalized perspective, however, there seems to still be some space for teleology, or for explanations of phenomena in terms of their aims and objectives rather than their causes (consequently, goal-oriented behaviors are more easily explained by their effects than their causes, but in the naturalized perspective, having a goal simply means showing a goal-oriented behavior).

Among others, Tegmark (2017, chap. 7), goes so far as to affirm that goal-oriented behaviors can be found in the laws of physics. In fact, based on the model of Fermat′s principle for the prediction of the behavior of light rays, these laws can be mathematically reformulated as optimization in relation to a quantity (for light, the minimization of travel time). However, we must not confuse the idea of a purpose and goal-orientation in the laws of physics with teleology in the world of life. At the basic level, the element that nature seeks to maximize is in fact entropy, i.e. total uniformity equivalent to thermal death, in accordance with the second principle of thermodynamics.

But the history of the universe is complicated by the presence of gravity, which counteracts the drive towards uniformity. And within thermodynamics itself there seems to be room for a dual domain with regard to teleology. On the one hand, there is a basic ″finalism″ inscribed in the laws of physics; on the other, there is the so-called dissipation-driven adaptive organization, proposed by England (2013). The idea is that random groups of particles tend to organize themselves in order to extract energy from the environment. In fact, dissipation is the process by which entropy increases, although its side effects may act in the opposite sense. If energy is converted into heat, entropy is produced, but at the same time a work is carried out that can reduce entropy within a delimited space. This is what self-organized systems do, which become more and more complex (and which one could think of as one of nature′s purposes).

Self-organized systems lead (in a way still substantially unknown to us) to life, and the most salient characteristic of living systems, as Schrödinger (1944) already observed, is that they maintain or diminish their entropy by increasing entropy in the surrounding environment. The second principle of thermodynamics does not allow for general exceptions, but at the local level there can be systems in which life increases its complexity by increasing disorder around it. In this way, paradoxically, life accelerates the general entropic process. A basic physical finalism can be seen in the particles' tendency to efficiently extract energy from the environment, and one way to increase efficiency for the particles is to make copies of themselves to have other energy extractors. Living forms are therefore organized systems able to replicate indefinitely.

The idea is that dissipation is the general ″purpose″ of the laws of nature and that replication, or the preservation of life - to use an anthropocentric terminology - is a sub-purpose, which often ends up taking over in the dynamics of the world of life. In this sense, evolution, if not guided by a true teleology, could have brought out processes and mechanisms, such as feelings and sensations in the human being, which often take the upper hand in the short term, producing behaviors that are not consistent with the goal of survival and reproduction. All this must be seen in the context of an unintentional explanation of the conduct of higher-order living beings.

According to Tegmark (2017. chap. 7), if by teleology we mean, as said, the explanation of things in terms of their ends rather than in terms of their causes, we can say that our universe is becoming more teleological. At first, all matter seemed destined for dissipation and global thermal death. With the emergence of life, part of matter began to focus on replication and its sub-purposes. And the living systems are reconfiguring more and more matter to support the realization of their own ends.

In this framework, a relevant role is played by the debate on the aims of artificial intelligence and of machines able to learn. The purposes of artifacts capable of autonomous action (even a trivial vacuum cleaner moving on its own around the living room) are programmed by the builders of the artifact itself. This also applies to more sophisticated examples of artificial intelligence. In this sense, the purposes of software and robots currently active in various fields are ″derivative″ or second-order purposes. But what is relevant here are the goals that could be autonomously developed by machines able to learn spontaneously thanks to computational skills and cognitive architectures far superior to human ones.

Consider a benevolent superintelligent AI - that is, programmed by constructors whose purpose is to make machines that are aligned to the goals of human beings: could it reconsider and modify these goals in a future stage, as a human being could? With regard to this we can only make speculations. But even today, AI is able to at least modulate its goals. And it can be hypothesized that a robot driven by advanced software would have fewer constraints than a human being, who is constantly conditioned by his or her genetic make-up and by the impulses it generates. But in what sense can an intelligent machine have purposes comparable or even superior to those of a human being? It is possible to speculate that even more powerful and self-learning artificial intelligences could find unencoded sub-purposes in relation to the general aims assigned to them, or that they may absolutize their own purposes by treating the rest as a means to achieve them.

On the one hand, for example, a series of algorithms could have the task of maximizing the happiness of a group of human beings. On the basis of this indication that we ourselves would consider too generic, each software could draw different concepts of well-being and ways of achieving it from the behaviors and beliefs expressed by the group under observation. They would thus seek to pursue different relevant goals, promoting spirituality or wealth, uniformity or diversity and so forth. On the other hand, a computer whose goal is to win at chess could use all the resources available for this purpose, including redirecting all the electricity supply of a city (including hospitals) to electronic memory elements to be added to its hardware. These purposes may be shareable or evil, but are they the expressions of a teleology comparable to the human one?

6. Husserl's Basic Teleology

In his last works and unpublished manuscripts, Edmund Husserl focused on the origin and trajectory of the self, in a natural and metaphysical teleological perspective (cf. Costa, 2009: ch 8). In the first area, Husserl saw the movement of life as a basic finalistic movement (not Darwinian, but such that can be read in terms of anthropic teleology). The nature of the self has its own path before becoming a personal subject: "The totality of human possibilities is present in the newborn child" (Husserl, 1973, p. 384). The ″I″ develops prior to the appearance of the intentional structure and of the I-manifestation-object triad, so that the transcendental emerges in successive phases.

As stated in Husserl (2006), a metaphysical will to life is expressed in every single human being. There is an instinctive intentionality manifesting itself as a drive that does not yet have a world of representations before it. The genesis of the subject and of experience coincides with the genesis of the kinesthetic system, i.e. a subject of will able to control their movements. The ″I″ stands as an ″I-center″ with spatial value, it is placed in a ″here″ that is a nucleus of possibilities for spatial movements. The living body comes to know itself as such in its ″here″. At the beginning there are automatic reflex-activity and organic functionality, so no action of the will is yet directed to a purpose (that is, one that is explicit and chosen by the subject).

The sense of purpose emerges little by little in connection to the will, together with the ability to represent. This is all based on kinesthesia, which for Husserl is an innate and instinctual - and therefore transcendental - aspect, and represents the subject's belonging to a life that precedes it (Husserl, MS K. III, 11 / 5a, quoted in Costa, 2009, p. 197). This is the basis of the formation of the world and of the perceptual horizon. To access this level the subject must be able to correlate and coordinate phenomenal representations and bodily movements. From the instinctual phase to the emergence of a purpose, the human being is traveling towards himself, so to speak, based on an intentionality and an innate direction that are manifested differently in each person. This presupposes a teleology of life that precedes consciousness. The world offers the not-yet-formed subject the directions to build her world.

For Husserl, there is an objective immanent teleology of the human being. But the transcendental will to life is a fact that does not allow for further explanation, if not, possibly, a reference to God. As individual transcendental subjects we are part of a universal dynamic of will to life (Husserl, 1973, p. 378). But this will to life is part of a rational process and of a rationality in the process of unfolding in which one can recognize a direction. In fact, for Husserl, the task of philosophical science, not always adequately realized in the course of its development, is that of recognizing and comprehending immanent teleology in the history of thought and in every single human action, as the presupposition and final outcome of any speculative inquiry. Following this perspective, Husserl proposes an ideal unity immanent to the history of philosophy, demonstrating how the search for an intimate truth and rationality of reality has characterized much of ancient and modern thought. The teleological idea is the tendency inherent in every human being in search of what satisfies the human ″metaphysical need″: it is the search for meaning and for the ″right path″.

Husserl describes a teleological movement of the world of life that seems to anticipate, at the level of philosophically informed intuitions, some elements of a naturalized teleology. For example, the role of kinesthetic competence in the encounter with the world recalls Gibson′s affordances (Gibson, 1979), the ″invitations″ of the world that the subject can grasp and lean on perceptually and materially, as if the world were tuned with life and life tuned with the world - something that can also be explained by the natural (and finalistic) mechanisms of biological intentionality seen above. The same universal dynamic of life can be read as an anti-dissipation tendency in Tegmark′s terms. In this sense one could then ask whether even an artifact, an intelligent machine, could take this path, albeit in a ″computational″ way. On the other hand, Husserl′s insights are framed within a more comprehensive teleology, which also includes culture and history understood as human events. It is therefore a wider teleology, metaphysically oriented beyond the narrow confines of a naturalistic explanation.

7. Is the Human Being Marked by a Specific Finalism?

Until recently, the specific difference of the human being has been sought in the comparison with other living beings, which was in some ways simpler. The specific difference, literally, is a characteristic that clearly shows the hiatus between the human and other species. In the thought experiment conceived by Jonas in his *Homo pictor* (Jonas, 1966, chap. 9), some astronauts disembark on an unknown inhabited planet and must evaluate if the native life form is comparable to human beings. Physical resemblance cannot be a discriminating factor, so another means of recognition is needed, one that is linked to action. Of course, talking about the search for an indicator of the life-form′s essence, as Jonas does, can be controversial today; however, the solution he offers is of great interest. It is about the *ability to represent.*

If the astronauts - says Jonas - went to a cave and found drawings with an optical resemblance to one of the life forms on the planet, they could say: ″This could have been made by ′humans′″. And the most crude and childish of drawings would be just as probative as Michelangelo′s art. What does representation prove? The fact that the author is capable of being ″symbolic″. But what qualifies the symbolic as unique? According to Jonas, a being that creates images is dedicated to the production of ″useless″ things, or has purposes that go beyond purely biological ones or, still, can pursue these purposes through ways other than the instrumental use of objects. In figurative representation, the object is appropriated in a new, non-practical way that bears witness to a new relationship with it.

In this sense, it is important to clarify what is meant by image and what are its characteristics. First of all, the representation must resemble another autonomous entity: this similarity must be produced intentionally, so as to be clear to the observer. This is because the key aspect seems to be that of incomplete resemblance. If you duplicate a thing, you get another example of the thing itself, or at least a ″photograph″ of it. This incompleteness - when not due to the inability of the author to produce a duplication - is what qualifies the image as such: it is the element of intentionality that guides the author, who makes a selection of the ″representative″ traits deemed ″characteristic″ of or ″significant″ for the object. As a consequence, underlines Jonas, a ″lack″ of completeness can mean something more in terms of essential resemblance. In some cases, the same (recognizable) intention of the author of the representation may take the place of the (evident) resemblance with the thing.

Given these properties of the image, one can ask what properties are required for a subject to produce and capture images. Producing an image presupposes the ability to perceive something as an image. And perceiving something as an image and not just as an object also means being able to produce one, without this implying some specific technical ability. For Jonas, perceiving similarity is a capacity for refined discrimination that belongs only to the human being as opposed to other animal species. In fact, the latter can grasp the representation of an object as equivalent to the object or as something other than the object (we see this difference at the level of external behavioral, and we do not know what it may correspond to in an animal′s internal state). But it is only the human being - speculates Jonas - that distinguishes the form represented from the vehicle of representation and knows how to grasp different degrees of similarity, something that has to do with abstraction, representation and symbolism.

Homo pictor is ultimately characterized by an eidetic control of the imagination, with its freedom of internal planning, which gives life to a freely chosen form, inwardly imagined and intentionally projected. The encounter with representation is therefore, for Jonas, the heuristic experience which manifests the ontological difference between the human being and other forms of life endowed with cognition and finalism. The freedom to produce images with resemblance to external reality without a purpose other than the purely aesthetic-recreational one seems to be the mark of the human.

If the *ability to represent* distinguishes the human being from other life forms, when it comes to smart machines things are certainly more complex. In fact, the ability to grasp similarities and differences and the ability to reproduce an object in an image with nuances of incompleteness and diversity are far more extensive in artificial intelligence than in human beings. The first capacity is simply the result of a greater computing power. The second ability can be borrowed from the style of an artist, which the intelligent machine is able to replicate on the basis of both guided and autonomous learning.

It can however be argued that one can nonetheless intuitively sense a difference between a human being and an intelligent machine. As mentioned, robots lack consciousness as a capacity to feel sensations; also - as some scholars claim - we cannot know if a software has states of a certain type, if it ″feels like″ anything to be an advanced robot. So, in our interactions with machines, we tend to favor intelligence over pure basic sensitivity. Which leads us back to our initial question: can smart machines be equated to human beings as subjects who have intentions and purposes?

The intuition of an essential difference seems linked, in the current state of things, to intentionality and teleology. Artificial intelligence can have the ability of representation, it can decide sub-purposes with respect to a general purpose, but does not seem (yet) able to autonomously act with a symbolic purpose, as defined by Jonas. The idea is that a person who creates images produces things that have no immediate use, or has purposes other than purely biological ones, or can pursue these purposes through ways other than the instrumental use of objects. The teleology of the human being is qualified by a directionality (the universal dynamic of Husserl′s will to life) but also by a freedom and a spontaneity endowed with meaning (i.e. not purely random). These properties are not (yet) found in intelligent machines.

I have used the word ″intuition″ in relation to the difference between humans and machines because the short investigation carried out so far seems to point towards a non-criterial idea of teleology. In other words, beyond the basic natural regularities implied by the laws of physics as we know them, it is difficult to identify a finalism devoid of exceptions (even the finalism of life is contradicted by many behaviors of the human being, such as suicide or more general non-adaptive choices). The non-criterial concept of specifically human teleology therefore has no precise necessary and sufficient conditions, but rather can be grasped by human beings themselves with a certain ″family resemblance″, to put it with Wittgenstein, in the world of life and in the inanimate world. *Homo pictor* has this quality, artificial intelligence does not (or at least not yet). *Homo pictor* can imagine new goals without being driven by some natural necessity. It can create a world of ideas and abstract representations, a world that brings us into a dimension of continuous phenomenological and cognitive growth - the world, as Husserl claimed, in which reason unfolds and in which the purpose is no longer only survival and reproduction, but some form of human flourishing in an indefinite sense.

This conclusion does not bring the desired clarity into the discussion, but perhaps helps understand the observer-relative character that the concept of teleology implies. This relative character does not belittle the ontological perspective, but rather accounts for an objective difficulty: one that intelligent machines themselves would probably find if asked about finalism and its manifestations.

REFERENCES

Bargh, J. (2017). *Before You Know It: The Unconscious Reasons We Do What We Do.* New York: Touchstone.

Brentano, F. (1884). *Psychologie vom empirischen Standpunkt,* Leipzig: Duncker&Humblot.

Costa, E. (2009). *Husserl.* Roma: Carocci.

Crane, T. (2001). *Elements of Mind: An Introduction to the Philosophy of Mind.* Oxford: Oxford University Press.

Dennett, D. C. (2017). *From Bach to Bacteria and Back: The Evolution of Minds.* New York: W. W. Norton & Company.

Descartes, R. (1641/1986). *Meditations on First Philosophy.* J. Cottingham, trans. Cambridge: Cambridge University Press.

England, J. L. (2013). Statistical physics of self-replication. *The Journal of Chemical Physics,* 139, 121923; https://doi.org/10.1063/1.4818538.

EU (2017). Report, with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL);http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//EN.

Frege, G. (1918). Der Gedanke. Eine logische Untersuchung. *Beiträge zur Philosophie des deutschen Idealismus, I,* 58-77.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception.* Boston: Houghton Mifflin.

Greenfield, A. (2017). *Radical Technologies: The Design of Everyday Life.* New York: Verso

Husserl, E. (1973). *Zur Phänomenologie der Intersubjektivität. Texte aus dem Nachlaß Dritter Teil: 1929-1935, Husserliana*, Bd XIV, I, Kern (Hrsg). Den Haag: Nijhoff.

Husserl, E. (2006). *Späte Texte über Zeitkonstitution (1929-1934): die C-Manuskripte, Husserliana*, Materialienbände, Bd. VIII, D. Lohmar (Hrsg). Dordrecht: Springer.

James, W. (1890). *Principles of Psychology.* New York: Dover, 1950.

Jaworski, W. (2011). *Philosophy of Mind. A Comprehensive Introduction.* Malden, MA: Wiley-Blackwell.

Jonas, H. (1966). *The Phenomenon of Life. Toward a Philosophical Biology.* New York: Harper&Row.

Kelemen, D. (1999). The scope of teleological thinking in preschool children. *Cognition*, 70, 241–272.

Kelemen, D., & Rosset, E. (2009). The Human Function Compunction: Teleological explanation in adults. *Cognition*, 111, 138-143.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional Physical Scientists Display Tenacious Teleological Tendencies: Purpose-Based Reasoning as a Cognitive Default. *Journal of Experimental Psychology: General*, 142(4), 1074-1083.

Kim, J. (2011^3). *Philosophy of Mind.* Boulder, CO: Westviewand Press.

Lavazza, A. (2015). *Filosofia della mente.* Brescia: Editrice La Scuola.

Millikan, R.G. (1984). Language, Thought, and Other Biological Categories. Cambridge, MA: The Mit Press.

Nagel, T. (1974). What Is It Like to Be a Bat?. *The Philosophical Review*, 83(4): 435–450.

Nagel, T. (2012). *Mind & Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False.* New York: Oxford University Press.

Nevejans, N. *et al.* (2018). Open letter to the European Commission: Artificial Intelligence and Robotics, http://www.robotics-openletter.eu.

Ryle, G. (1949). *The Concept of Mind.* London: Hutchinson.

Schrödinger, E. (1944). *What Is Life? The Physical Aspect of the Living Cell - Mind and Matter.* Cambridge: Cambridge University Press.

Searle, J.R. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3, 417-457.

Strawson, G. (1994). *Mental Reality.* Cambridge, MA: The MIT Press.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence.* New York: Knopf.