Free will and algorithms: a typical androrithm

Cristiano Calì

University of Turin - Institute for Ethics and Emerging Technologies
cristianocali30@pust.it

Abstract

This contribution moves in the specific area of the philosophy of mind and, in particular, in that of the philosophy of free will. The question of free will, in fact, has always been at the center of philosophical debates and is still an open question today. The aim of this paper is to use the discipline of artificial intelligence as a magnifying glass for the free will problem in order to identify, through it, how this cognitive capacity is an *androrithm*: an element specific to the human being and irreproducible. Through an analysis of the similarities and dissimilarities that the question of artificial intelligence and that of free will share, and a brief review of the various types of freedom that - in the face of contemporary debate - could be present in both human beings and machines, we will come to the conclusion that the so-called ambitious free will, if it exists at all, can never be reproduced and is therefore characterized as a constitutive element of the human being.

Introduction

For more than fifty years now, we have been witnessing remarkable developments in the field of what in 1956 was called – raising countless critics – *artificial intelligence* (AI). The advent of AI has also been an extraordinary revolution for philosophy, especially from an epistemological point of view, since it has prompted philosophers, and especially philosophers of the mind, to shift the center of gravity of their interests from conceptual, purely philosophical analysis to the study of the human being (including his brain) as an animal among other products of biological evolution (Nannini, 2011, p. 181). Most remarkably, the last two decades have seen the realization that artificial intelligence is an extraordinary opportunity and a formidable perspective lens for rethinking anthropology (Krienke 2020). It is my intention, then, to see the intrinsic affinities of the relationship between free will and artificial intelligence in order to try to understand whether the irreproducibility in machines of freedom can also be followed by its irreducibility. In particular, I will try to show how the cognitive capacity that is called free will is a typical *androrithm*, because it is irreproducible in the machine and, therefore, is a an element that can be defined as an *elementum constitutivum* (constitutive element) of the human beings.

1. Free will and artificial intelligence in the mirror

The similarities between free will and AI can be traced as far back as their definition. For both notions, in fact, there is considerable difficulty in providing a correct and exhaustive definition that is unanimously agreed upon. I like to resort to a very incisive expression used in an interview by Luciano Floridi:

> Friendship, AI and many other things in life are like pornography: they are not definable (in the strict sense in which water is definable and defined as H2O) but we recognise them when we encounter them (Krienke, 2022, p. 41).

In this list of indefinable things, I believe that freedom can also be included, which is complex to characterize unambiguously but nevertheless we recognize it when we see it, or rather, when we experience it. To be fair, it should be pointed out that it is not correct to say that "AI cannot be defined by listing necessary and sufficient conditions" (*ibid*., 42), but rather, that there is no common agreement as to what the necessary and sufficient conditions for AI are to arise, just as there is no agreement whatsoever as to what the necessary and sufficient conditions for freedom are to arise.

A second affinity between artificial intelligence and free will is given by the fact that in both cases a dualist ontology, however problematic, would seem to be the best context in which to situate both self-determination and strong AI. In the case of artificial intelligence, in fact, materialism is anything but fundamental to the *Artificial General Intelligence* (AGI) project, since for a materialist, what counts is physics, the stuff of which something is made; for the proponents of strong AI, on the other hand – paradoxical as it may seem – the basic structure cannot and should not constitute a discrimen: what really counts is only functionality.

A third aspect, which instead brings out a dissimilarity, is constituted by the notion of omission. Omission is not contemplated for machines; indeed, it is commonly assumed that the distinction between a human being and a machine is the unpredictability of the former and the regularity of the latter. The human being, on the other hand, has always been granted room for manoeuvre in which he can act with freedom – or it would be better to say not to subscribe to any of the previously analyzed accounts – and with creativity. The human being is given to act or not to act at all, therefore also to "make" an omission. I think this may be a relevant element.

For an engineer, in fact, if the machine, being in a certain state, does not choose, it does not move on to the next state[1] (and this means that that mechanism has crashed); the human being, on the other hand, can "make an omission", without thereby returning a pathological malfunction.

Conversely, an aspect inherent to freedom that has always led to humans being equated with automatons is related to determinism. The notion of determinism is fundamental to AI since an algorithm is defined as a procedure that solves a given problem through a finite number of elementary, clear and unambiguous steps (Longo and Scorza, 2020, p. 8). It is no coincidence that any computer could be taken as a perfect model of both Laplace's ideal and Democritus' ideal: even if time were turned backwards, the computer would always act in exactly the same way, always taking exactly the same number of steps to solve the problem. A comprehensive description of how determinism is binding for humans and machines is provided by Ted Honderich:

> Determinism is only a view of our nature-in essence, the view that ordinary causation is true of us and our lives, that in our choosing and deciding we are subject to causal laws. In this use of the word, determinism comes to no more than a yes answer to the question of whether we are in one fundamental way like plants or machines. Determinism in this sense does not include or imply an answer to the question of whether we are free or not. That question, maybe surprisingly, is left pretty well untouched (1993, p. 22).

Laplace's genius is also useful to recall another analogy between AI and freedom that revolves around the concept of predictability. The Laplacian dream, in fact, may seem closer today than it once did, thanks to the highly efficient predictive capacity of AI: a field in which machines far surpass humans and which is always growing exponentially. We live in a world so full of data that predictive capacity is almost indispensable (even in the context of scientific research). We must ask ourselves: what would our actions be like if a machine provided us with a highly probable prediction of future reality? Here again, the picture is not futuristic but perfectly concrete, since several algorithms today have a predictive capacity, based on a very high computing power, that is far superior to that of the human being. The ethical and social implications, when faced with a scenario in which machines can perfectly predict a person's growth, his developments, his actions, are particularly significant; nevertheless, it is precisely this astonishing predictive capacity of AIs that seems to collide with any theory espoused in reference to freedom, and not only the most beaten-down ones (such as compatibilism and libertarianism), but also the more abstruse ones (such as fatalism or theological determinism). To explain this irreducibility, I refer to a 2015 Steven Spielberg film entitled *Minority Report*, based on Philipp Dick's extraordinary novel of the same name. The film is set in the Washington of 2054, a city in which murders have been eradicated thanks to the predictions of three individuals with extrasensory powers who are able to foresee the crime, put the police on notice, who then manage to foil the murder (Guzzonato, 2022).

The film is based on the assumption that the world is perfectly deterministic (otherwise prediction would be impossible) and in the film, indeed, everything seems predetermined. This total predetermination is not disproved even when the protagonist, played by Tom Cruise, who was predicted to kill a man, tries to change reality but eventually (accidentally) fires the fatal shot, thus proving that the prediction was correct. The same predetermination is also found in one of the final scenes when the police chief, played by Max von Sydow, kills himself in order not to allow the prediction system to be disproved. Everything seems to be the perfect realization of the deterministic dream, except for the consideration that perhaps nothing that happens was determined. The three individuals with extrasensory powers, in fact, did not see the real future but one of the possible courses of action (not the only possible course of action, as determinism would have it) and this because the police, in fact, interrupted the course of action leading to the murder. The police, in fact, foiled the murders before they were committed and arrested the potential perpetrators. In other words: the DC system of the future does not punish the deed (which does not happen) but the intention to commit it, an intention that never translates into deed. The three individuals ultimately foresaw both (attempted) murder and arrest and thus, in fact, that grisly future never materialized. Such a prediction, however, has nothing to do with Laplace's deterministic genius but with something entirely different: a knowledge so high – what the Spanish Jesuit Luis de Molina called scientia media in the 1500s – that it allows God to know the future continents. The curious thing is that the notion of scientia media was introduced by Molina to circumvent the problem of determinism, not to corroborate it.

Such an account is useful to show how even the remarkable predictive capabilities of an algorithm would not be able to perfectly predict the future. More, even if we were in a deterministic context, the understanding of freedom as defended in the philosophy of mind by agent causation theorists would not allow a super-intelligence to predict its own actions: freedom understood as creativity would always escape even this possibility of prediction (and perhaps this is precisely what Dick's account meant).

2. Freedom in machines?

Nevertheless, it is possible to renounce the ideal of predictability and safeguard determinism at the same time. In this second scenario, in fact, if one did not claim to defend free will but only a form of compatibilist freedom, it would be conceivable to understand this compatibilist freedom as a reproducible freedom (assuming one arrived at strong AI). Kant had already exposed this issue:

> If a human being's actions insofar as they belong to his determinations in time were not merely determinations of

[1] When I speak of omission for algorithms, I mean the choice of doing nothing while receiving a specific input. This perspective in the Middle Ages was called the theory of pure omission.

> him as appearance but as a thing in itself, freedom could not be saved. A human being would be a marionette or an automaton, like Vaucason's, built and wound up by the supreme artist; selfconsciousness would indeed make him a thinking automaton, but the consciousness of his own spontaneity, if taken for freedom, would be mere delusion inasmuch as it deserves to be called freedom only comparatively, because the proximate determining causes of its motion and a long series of their determining causes are indeed internal but the last and highest is found entirely in an alien hand. (Kant, 2015, p. 82).

Embracing a purely mechanical freedom would, in fact, be tantamount to being nothing but robogeeks, creatures who are somehow disposed to cast away the very essence of their humanity and embrace a personal identity as automatons (Wegner, 2002, p. 43). Entities of this kind, however, would not be machines that make free decisions, but – recovering the distinctions proposed in chapter one – only machines that operate automatic decisions or, with a more exact wording, automatisms. These kinds of "actions", I believe, already exist: it is now well known that AI algorithms incorporate unprocessed information, raw data, which can then be analyzed, and on the basis of which artificial intelligences will be able to develop their own implicit knowledge. In the light of that data and accumulated experience, the machines act through a process that is referred to in psychology as an automatic process. This process, however, is not similar to what we call weighting but rather is similar, if not identical, to what happens when, while we are driving, we see a child jaywalking and brake. In such cases we do not proceed to a perfectly conscious rational reasoning but make the choice solely on the basis of our experience and brake. If the freedom we want to reproduce is a mechanical freedom, I think it is only a matter of time: we only need to know perfectly how the brain works to have its counterpart on a silicon basis. Even if we want to support the reproducibility of this modest form of freedom, a corollary of it (and not of the ambitious freedom) poses additional problems.

Compatibilist freedom, in fact, is that same freedom defended by Daniel Wegner: we have the feeling of acting, we feel acts as our own, but in truth we have no control over them.

> The unique human convenience of conscious thoughts that preview our actions gives us the privilege of feeling we willfully cause what we do. In fact, however, unconscious and inscrutable mechanisms create both conscious thought about action and the action, and also produce the sense of will we experience by perceiving the thought as cause of the action. So, while our thoughts may have deep, important, and unconscious causal connections to our actions, the experience of conscious will arises from a process that interprets these connections, not from the connections themselves. (*ibid*., p. 125).

Assuming for a moment that this account is true, we should say that this dimension is also irreducible and irreproducible. The difficulty arises, in fact, not in reference to having access to unconscious thoughts – which by their very definition cannot be accessed consciously – but to our sensation of will (which would be the result of an interpretation of such thoughts). The problem lies, firstly, in the fact that much less is known about these processes today than is known about the human brain as a whole; secondly – and this is the most remarkable fact – It lies in the fact that precisely this interpretation of the mechanisms in agential terms is inaccessible to us (otherwise it would not be an illusion), and by virtue of this it would constitute us as agent beings. In this regard, it was Huxley who suggested that there will always be a difference between humans and machines because we are "conscious automatons" (in *ibid*., 2019, p. 46). The problem therefore lies in the fact that in order to have a strong AI, an AI that is equal if not superior to the human being, we would have to reproduce something that is not only inaccessible but that, should it become accessible, would cause us to lose our very essence as human beings, as conscious beings.

If, therefore, it is possible to reproduce a modest, or compatibilist, freedom, it is equally possible to reproduce an indeterministic (I deliberately do not say *incompatibilist*) freedom, that is, the freedom that critics of libertarianism claim is defended by radical and causal libertarians: a freedom that coincides with chance. In order to have a machine that acts according to chance, we don't even need to resort to those extraordinary artificers of computing that are now known as quantum computers. Today, in fact, even perfectly deterministic AIs such as satellite navigators are programmed so that from time to time they can act stochastically, completely randomly, to find, for example, new roads for the route indicated by the user. True, some may say that the algorithm is programmed to act that way anyway, but in fact its action would not be predictable (as it is random): we might at best opt to describe this form of "freedom" as "free necessity".

Whether we have deterministic or stochastic systems, however, we ultimately do not have freedom, at least not in the sense I have defended in the preceding paragraphs. To summarise this first point I quote Wegner:

> Imagine […] a person in which there is installed a small unit called the Free Willer. This is not the usual psychological motor, the bundle of thoughts or motives or emotions or neurons or genes – Instead, it is a black box that just does things. Many kinds of human abilities and tendencies can be modeled in artificially intelligent systems, after all, and it seems on principle that we should be able to design at least the rudiments of a psychological process that has the property of freely willing actions. But what exactly do we install? If we put in a module that creates actions out of any sort of past experiences or memories, that fashions choices from habits or attitudes or inherited tendencies, we don't get freedom – we get determinism. The Free Willer must be a mechanism that is unresponsive to any past influence. (*ibid*., p. 322).

What emerges here is a confusion between free will and *libertas indifferentiae* that has often recurred in the debate on free will. However, it is useful to ask whether *libertas*

indifferentiae is at least reproducible. While it has been described in many quarters as impossible for human beings, it may well be possible for machines. If it is true, in fact, that complete passivity will always be impossible for a living being, so that it will be totally indifferent to perform *a* or *b*, for the machine this would seem to be more feasible. More recent decision-making algorithms, on the other hand, have shown multiple cognitive biases and it has had to be recognized how difficult it is to have an unbiased, indifferent algorithm. Adina L. Roskies says:

> If we have no access to any proposition relevant to our choices, then we really do seem to be autonomous people who do not act on reasons, but simply behave in a way that is externally describable in terms of reasons (2019, p. 66).

And yet, on closer inspection, not only do we fail to act on the basis of reasons - even the use of will itself - but apparently neither do AIs manage to do so (where by reasons in this case we mean the data provided to the algorithm).

To make this even clearer, let us look at the case of IBM's now famous computer, *Deep Blue*, which in 1996 defeated the then world champion in the game of chess. The computer emerged the winner of the game because thanks to its computing power, it made all the exact moves. Now in performing these actions, the computer was not free since it was determined to make those moves with a view to victory. Take the case above: let us imagine that the computer has been instructed to make one random move every twelve exact moves. Will that thirteenth move be free? Again, the answer would be negative because Deep Blue's behavior - however determined – would be guided by a random criterion. A third scenario, however, raises more perplexity: would the computer be able – every 3, 6, or 7 moves (I can't say *when it wants* but my intent is understood) – to make a wrong move? Mind you, I am no longer saying a random move (since a random move might turn out to be a right one) but a wrong one. Let me add a detail: although it is particularly difficult for a computer to make a wrong move because of, say, a computational error or because it is faced with another, better-performing computer, could a computer make a deliberately wrong move in the same way that an award-winning world champion would make a mistake because, say, he is tired and is making a mistake in order for a young up-and-coming talent to win? The answer is not so simple, and I think the question itself needs to be rephrased in the following terms.

What kind of freedom is reproducible in order to have AIs that are more and more like human beings? In my humble opinion, none, and not because the skills of engineers are inadequate and so it would only be a matter of patience, but because we simply do not know what freedom is and when we try to grasp it, it appears irreducible to our understanding. I remember the words of Professor Luca Gambardella, president for many years of the *Dalla Molle Institute for the Artificial Intelligence* in Lugan (Switzerland) – one of the world's first centers for AI research – who told me: "Tell me what something looks like and I will reproduce it for you. It will be a matter of time, we will involve hundreds or thousands of engineers, but we will reproduce it". Here is the real problem: we - in the face of freedom - do not know what we have to reproduce. Paolo Legrenzi and Carlo A. Umiltà, in criticizing the localizationist principle in neuroscience, had already posed this dilemma: It is our conviction that first it is necessary to ask "what we are localizing" (2014, p. 90). This question, as researchers strive to build a strong AI that is sentient and free, should be rephrased as follows: "What are we reproducing? What to reproduce, however, still remains, after several centuries, a mystery in the above sense, something that exceeds not only our knowledge but also our own reality.

3. *Agere sequitur tantummodo esse* (acting is an exclusive prerogative of being)

If it is not possible to know something, is it possible to reproduce it? Probably, the answer is negative. With such a statement, I do not wish, once again, to endorse any sceptical perspective. Instead, through the reference to AI, I have tried to suggest a final rethink in favor of the question of freedom. In this I have followed the advice of Edsger Dijkstra, professor of computer science at the *University of Austin* in Texas, for whom "computers are only useful in making us understand what they cannot do" (Rossi, 2019, p. 69).

While it is true that freedom cannot be reproduced because we do not yet know it thoroughly, should we come to know and understand freedom, it could also be reproduced in a sophisticated AI. What misleads us is precisely the fact that the brain turns out to be an extraordinary computational machine.

> Curiously, pure intellectual processes lend themselves well to an algorithmic account and do not appear to be dependent on the substrate. This is the reason why well-conceived AI programs can beat chess champions, excel at Go, and drive cars successfully. However, there is no evidence to date to suggest that intellectual processes alone can constitute the basis for what makes us distinctly human. (Damasio, 2018, pp. 302).

The freedom that our intuition gives us, on the other hand, is an ambitious freedom, and although in my view it is in principle knowable, it would still escape reproducibility because, as a creative vis emerging from the processes of the entire living organism, it would not be reducible to its functions or internal mechanisms alone. In order to criticize this perspective, which, in my view, can only be loosely defined as libertarian, we have often resorted to the expression that became famous with Gilbert Ryle of "ghost in the machine", since a free will capable of originating actions is believed to be like a ghost in the control center of our brain. Thus, states Damasio:

> Saying that living organisms are algorithms is in the very least misleading and in strict terms false. Algorithms are formulas, recipes, enumerations of steps in the construction of a particular result. Living organisms, including human organisms, are constructed according to algorithms and use

> algorithms to operate their genetic machinery. But they are not algorithms themselves. Living organisms are consequences of the engagement of algorithms and exhibit properties that might or might not have been specified in the algorithms that directed their construction. Most important, living organisms are collections of tissues, organs, and systems within which every component cell is a vulnerable living entity made of proteins, lipids, and sugars. They are not lines of code; they are palpable stuff. (*ibid*., 448).

This judgement is also shared by profound connoisseurs of AI:

> Perhaps, we must realise that intelligence is not the trait that most characterises human beings. After all, animals, especially primates, also display amazing cognitive abilities, but only our species possesses the desire to create (Turi et al., 2019, p. 14).

This desire to create cannot and must not be understood trivially as a reproductive or productive desire (in the sense of the Greek τέχνη) but as a desire to be creators, to be able to create something that 'belongs to the subject' in an inalienable way: the production of actions that are exclusive to the acting subject and only to him. The desire to create is also accompanied by the sensation of being creators, a sensation of origination that, with the goodwill of many neuroscientists, is also inalienable[2]. One would then have to disprove the thesis defended by Harari when, recognizing the fields in which AI excelled by far over human beings, he stated:

> We don't know of any third field of activity – beyond the physical and the cognitive – where humans will always retain a secure edge. (Harari, 2005, p. 44).

Perhaps this field of activity exists and is both physical and cognitive (also because the two should not be separated). I have tried to suggest just that by referring to AI. Freedom, rather than one of the many cognitive capacities, is configured – to use a term of Gerd Leonhard and taken up by Federico Cabitza – as an *androrithm*, i.e., something that we can consider irreducible to algorithmic conversion, i.e., to coding and symbolic representation (Cabitza, 2021, p. 54). Most probably, it is neither intelligence nor conscience that are the proprium *constitutivum* (specific element) of the human being, but freedom, not because it is something divine, but because it is something purely and totally human, which is constituted at one and the same time as a limit but also as a yearning.

[2] Speaking of the creation of action, the meaning must be that of being the originator of one's own action. The latter - for the theory of agent causality - is the human being's own possibility of initiating new causal chains. Understood in this way (and if the theory of agent causality is proved) free will is a specific element of the human being and for these reasons the freedom we can say of algorithms, like ChatGPT, will be something totally different, because that kind of freedom will only be a deterministic or stochastic freedom and not a creative freedom.

References

Nannini S., (2022). *L'anima e il corpo. Un'introduzione storica alla filosofia della mente*, Laterza, Roma-Bari.

Krienke M., (2020). I robot distinguono tra bene e male? Aspetti etici dell'intelligenza artificiale, *Aggiornamenti Sociali*, 4, pp. 315-321.

Krienke M., (2020). Svolta nell'antropologia? Su rischi e opportunità dell'IA per la filosofia, Anthropologica, 1, pp. 67-81;

Cabitza F., (2021). *Deus in machina? L'uso umano delle nuove macchine, tra dipendenza e responsabilità*, in Id. and Floridi L. (2021). *Intelligenza artificiale. L'uso delle nuove macchine*, Bompiani, Milano, pp. 9-111.

Harari Y.N., (2005). *21 Lessons for the 21st Century*, Penguin, New York.

Turi N., Gori M. and Landi M. (2019). *Guida per umani all'intelligenza artificiale. Noi al centro di un mondo nuovo*, Giunti, Firenze-Milano.

Damasio A.R. (2018). *The Strange Order of Things: Life, Feeling, and the Making of Cultures*, Pantheon Books, New York.

Legrenzi P. and Umiltà C. (2014). *Perché abbiamo bisogno dell'anima. Cervello e dualismo mente-corpo*, il Mulino, Bologna.

Rossi F. (2019). *Il confine del futuro. Possiamo fidarci dell'intelligenza artificiale?*, Feltrinelli, Milano.

Kant I. *Kritik der praktischen Vernunft*, 1788, (eng. ed. *Critique of Practical Reason*, ed. by M. Gregor, Oxford Univrsity Press, Oxford-New York 2015) .

Roskies A. (2019). *Esiste la libertà se decidono i nostri neuroni?*, in De Caro M., Lavazza A. and Sartori G., editors, (2019) *Siamo davvero liberi? Le neuroscienze e il mistero del libero arbitrio*, Codice, Torino, pp. 51-69;

Wegner D.M. (2019), *L'illusione della volontà cosciente*, in De Caro M., Lavazza A. and Sartori G., editors, (2019) *Siamo davvero liberi? Le neuroscienze e il mistero del libero arbitrio*, Codice, Torino, pp. 21-49.;

Wegner D.M. (2002). *The Illusion of Conscious Will*, MIT Press, Cambridge/MA 2002.

Guzzonato C. (2022). *Come in Minority report con Tom Cruise: l'IA prevede I crimini negli Usa*, in https://www.focus.it/tecnologia/innovazione/IA-intelligenza-artificiale-usa-prevede-crimini-rivela-pregiudizi (cons. 18/7/2022).

Honderich T. (1993). *How Free Are You?*, Oxford University Press, Oxford-New York 1993.

Longo A. and Scorza G. (2020). *Intelligenza artificiale. L'impatto sulle nostre vite, diritti e libertà*, Mondadori, Milano.