PENROSE AND HIS POSITION AGAINST THE POSSIBILITY OF THE COMPUTABILITY OF THE HUMAN MIND AND CONSCIOUSNESS

DANIEL HEREDIA
Universidad de Sevilla

1. INTRODUCTION

Many times, and more so these days, society has come to wonder if computers will ever possess intelligence or experience situations similar to those of humans. A suggestive response on this topic is given by the physicist-mathematician, Nobel Prize winner in Physics (2021), Roger Penrose, who is against this possibility becoming a reality. Penrose does not deny that a computer acquires some qualities that make it resemble a human being. In fact, recognize it very willingly (Penrose, 2012: 60-61), since the advances of Artificial Intelligence are undeniable.

But the underlying question is another. And which one is this? Far from wanting to make a philosophy of language or analysis, Penrose finds the crux of the matter in the expressions "intelligence" and "experiencing" of the initial question. In order to start the debate about the possibility of artificial intelligence or deny it, it is necessary to be clear about what is meant by intelligence. Is a computer intelligent, capable of playing chess at the level of a professional chess player or performing mathematical calculations at a speed that no human being can reach? The superiority of any chess module over any Grandmaster, or the possible calculations that any computer can make when faced with problems of a certain complexity show capabilities that are worthy of being considered intelligent. But it seems that the concept "intelligence" necessarily leads us to go further.

The structure of this work is composed of various sections, which contain the following topics.

In section 2 we will come into contact with the fundamental concepts within the debate on the possibility of artificial intelligence. Such terms as "intelligence", "knowledge" or "consciousness" will be treated, through the studies of Gardner and Penrose himself. Section 3 is devoted to trying to get an overview of Penrose's defense in this debate. Section 4 is divided into two subsections. In 4.1., we will see the points of view within the debate that Penrose takes into account. And in 4.2., we will have the opportunity to expand the number of points of view, especially thanks to Sloman's criticism. And in the last section 5 we will see an illustrative example of what Penrose tries to defend in the debate.

2. GOALS: ACLARATION OF CONCEPT INTELLIGENCE, ITS TYPES AND THE POSSIBILITY OF AN ARTIFICIAL INTELLIGENCE

Intelligence[55] is commonly defined as the mind's ability to think, learn and make decisions. In the first instance it seems that intelligence can perfectly be what we are told with this definition. But a second analysis alerts us that this conception is missing some more ingredients. Examples of them are the logical, creative, orientation, self-awareness, understanding of feelings, memory or even the ability to teach. All of these capabilities, without a doubt, require intelligence or, rather, are constitutively intelligent. But, which of them defines intelligence? Is there a hierarchy between these capabilities? Currently there are trends in which raising these types of questions makes no sense.

One of the most important is the one that follows the theory of multiple intelligences, proposed by Howard Gardner. According to Gardner,

[55] It is worth mentioning that in the vast majority of cases of my search in encyclopedias, books, articles and research papers about the concept of intelligence, the results obtained were related or directly focused on Artificial Intelligence. In this way it is easy to understand the state of research interest in this concept and under what criteria it is usually constructed. However, I have thought that it is best to give a general definition (or rather a definition of human intelligence) and then see how it fits with the doctrine that almost completely dominates it.

intelligence should not be understood as the balanced development between different types of intelligence. And what are those types of intelligence? He classifies them as follows: musical intelligence, kinesthetic-bodily intelligence, logical-mathematical intelligence, linguistic intelligence, spatial intelligence, interpersonal intelligence, intrapersonal intelligence and naturalist intelligence.

The different types of intelligence must be understood as independent. For example, the fact that someone is not very socially skilled does not exempt him(her) from being intelligent if they have a great capacity for mathematical-logic. In fact, it is common to excel in one intelligence, instead of having great ability in all of them. However, the independence of the different intelligences cannot be complete. There are rarely cases in which only one intelligence is possessed or that the different intelligences do not have any type of relationship. There are intelligences that feed each other. Gardner calls this feedback *combination* (Gardner, 2006: 22).

This, without a doubt, is the main characteristic of the theory proposed by Gardner, or, at least, the one that raised the most controversy and continues to raise in debates concerning intelligence. But before moving on to see the reason for this controversy, let's look at the other two aspects that stand out as fundamental, since these also contribute to a more extensive definition of the main concept.

The first of the characteristics is the uniqueness of the intellectual profiles, due to their material composition. Two people (or three or four...) cannot possess the same intellectual qualities, since this ultimately depends on the material that makes up their bodies (which is always different!). Gardner uses the example of identical twins. Although the appearance of such individuals is similar, the material from which they are made is different and particular. Therefore, it is understandable that they may have completely opposite intellectual abilities.

The second has to do with the importance of the ability to choose, or rather to choose well (or as we have been seeing it until now, the ability to solve problems). This aspect is very important within Gardner's

scheme, to the point that in all types of intelligence the ability to solve problems is implicit to different degrees.

Let's see what criticism the theory of multiple intelligences receives. This will help us better understand how the term intelligence is currently used, which is what we should be clear about.

Psychologists opposed to Gardner agree that intelligence is made up of different aspects. But these aspects are independent in such a way that someone can possess a high degree of one of them and be null in another, all without losing the condition of being intelligent, is necessarily incorrect. If intelligence is characterized by something, it is by having a balance between different abilities. That is why the information that intelligence tests offer us is so important for this side of the debate. But, to what extent is it convenient to trust to these tests? Not even the supporters of the development and improvement of these intelligence tests demand such a bold goal. What they do seem to be unwilling to give up is experimental data and empirical evidence, which is precisely what theories of multiple intelligences lack[56].

The underlying question is whether intelligence depends on a balance between its different aspects, with all of them having to be related, or if, on the contrary, there is independence between them in such a way that we can speak of multiple intelligences. The current trend (and let us understand that a real consensus as such does not exist) is to understand intelligence as that set of different abilities that are related to each other and that help solve problems. The more intelligence, the easier it is to solve problems. That is to say, a kind of combination between the two positions seen is what is generally accepted. However, it seems that elements are still missing, at least conceptually. In neither of the two theories are concepts as important when talking about intelligence as understanding or consciousness considered. And I am not suggesting that we ignore them. In my opinion, they make a more serious mistake, that is, taking such concepts as accepted or understood.

[56] This is the main accusation of the detractors of the theory of multiple intelligences. One of the greatest exponents of this position was the psychologist Hans Eysenck, who fiercely criticized Gardner in particular.

Later we will see what place these conceptions occupy with respect to that of intelligence. Since for now it is not conflictive to understand intelligence in Gardner's way and his detractors, we continue with the exposition.

Next, we will see how we can relate such a difficult concept to the possibility of artificial intelligence.

Given what we have seen, we can ask ourselves: is artificial intelligence possible? As people related to the field of philosophy, we can ask ourselves any questions we want (of course!). Another issue is getting a definitive response. It is advisable to continue clarifying the different concepts that will be seen throughout it and with the greatest of caution.

Until now we have seen the expression Artificial Intelligence on several occasions. It is time to determine which definition corresponds to know exactly what type of Artificial Intelligence we will encounter from now on.

Artificial Intelligence is a philosophical doctrine that defends the possibility of creating an intelligent entity through different processes and based on a device[57]. According to Russell and Norvig (2003), the ambition of the supporters of this doctrine is divided into two aspects:

1) Those who seek to achieve the way of acting and thinking of human beings, so it is strictly necessary to get to know our nature.

2) Those who seek to achieve the way of acting and thinking typical of rationality, requiring knowledge of the perfect functioning of reason, with the ambition to surpass human reason.

From now on, to distinguish them more schematically, I will refer to the first aspect as AI-1 and the second as AI-2.

The AI-2 aspect does not necessarily have to aspire to surpass the human being by knowing his nature. In fact, to be considered a doctrine other than AI-1, it must be understood in such a way that it aims to achieve an intelligence different from that of the human being. In AI-1 it is assumed that just when machines reach the level of human thought they will be superior. The main problem is whether this reach by

[57] Generally said device is electronic.

machines will take place or not. This is not really a problem for AI-1 supporters, who are convinced that it will happen.

The difference between AI-1 and AI-2 is that for AI-2 it is not necessary to reach the level of human thought. We must go beyond this, perfecting the path towards correct rationality. When AI-2 defends the obtaining of said rationality, it does not mean that it considers human beings as lacking reason. It is not about taking away from human beings a characteristic that has always been associated with us. For AI-2, our reason is not perfect and it should not be impossible for a machine to be able to surpass the human being in this aspect. On the other hand, this side is also aware of the difficulty of the problem. But, as in AI-1, the goal is considered achievable.

Another common trait that exists between AI-1 and AI-2 is that both aim to reach their respective goals (which, as we have seen, is actually the same) through computing[58]. But what is computing? It is generally understood as the action of introducing a set of scientific knowledge and methods into a machine so that it can handle them automatically. As Penrose clarifies (Penrose, 2012: 32-34), computing consists, to a greater extent, of two types of procedures:

a) top-down, which are those whose structure is fixed, allowing a unique and, therefore, precise solution. And on the other hand we have the procedures:

b) bottom-up, which are understood in such a way that their rules may vary. The solution may also be precise, but it is not a necessary step within its structure.

But although computation is a common aspect in both doctrines, it also contains the fundamental difference seen above. While in AI-1 the computation must be based on the way a human thinks and acts, for AI-2 this may even be irrelevant. This aspect is more conflicting for AI-1 than for AI-2, in the sense that AI-1 is assuming that human intelligence ultimately responds to a computational activity. And this is by no means

[58] In fact, there are those who understand Artificial Intelligence as Computational Intelligence. See, for example, Poole, Mackworth and Goebel (1998).

obvious. In fact, as we have seen above, the debate that arises from this idea will be the one on which this work will be focused.

Where is Penrose's position on this whole matter? Penrose understands the concept of intelligence as we have seen it, but with the particularity that intelligence is subordinate to consciousness[59]. Does he mean that intelligence is irrelevant or at least minimally important? Not at all. As its name indicates, the debate we have been seeing is centered on the possibility of an artificial intelligence. Of course, this conception is important to Penrose. The point is that he considers it essential to understand this – let's call it – hierarchy between intelligence and consciousness, since it contains the key to everything. For Penrose, it is impossible for machines to have human-like intelligence because they cannot become conscious:

> […] In my own way of looking at things, the question of intelligence is a subsidiary one to that of consciousness. I do not think that I would believe that true intelligence could be actually present unless accompanied by consciousness (Penrose, 1991: 505).

That is to say, the possibility of machines possessing intelligence is not impossible. In fact, he may consider it as a more than probable possibility. But everything changes when the goal is to reach human (conscious) intelligence, and this is something that he flatly denies. At first, then, Penrose does not launch his criticism against those who follow the AI-2 side, but against the supporters of AI-1.

This last idea is of great importance in Penrose's thought, but it does not offer us a general plan of the Penrosean scheme. I have reserved this purpose for the next section.

---

[59] Penrose finds it necessary to adequately understand four fundamental concepts in this debate. These are “awareness”, “understanding”, “consciousness” and “intelligence”. For him, and we will understand it in the same way from now on there is a relationship between these four concepts. That relationship is one of subordination or requirement. Penrose explains it like this:

“intelligence” requires “understanding” and “understanding” requires “awareness” (Penrose, 2012:54).

3. DISCUSSION: OVERVIEW OF PENROSE´S VISION

That machines are capable of displaying some human abilities and even several of them at the same time does not mean that they can possess the nature of our species, since this is too complex to be reduced to such a conception. This is Penrose's answer to the question of whether human-like artificial intelligence is possible. But is our nature really unattainable? What makes us so unique that it is impossible to make our abilities computable? Perhaps, precisely, non-computability? Penrose believes that the answer to this last question lies the crux of the matter. In human nature there are non-computable traits, traits that characterize us with respect to the most sophisticated machines that human beings can create. This is an argument demonstrable through science.

Despite offering an overwhelming response, the truth is that Penrose appeals to a project-based response[60]. At present no definitive verdict can be obtained for the debate of the computability of the human mind and consciousness. But in the future nothing is ruled out.

Penrose postulates that it is essential to understand quantum mechanics to be able to shed light on the true functioning of things (including our brains) and to be able to observe that computing our consciousness is impossible. But this is obviously not an easy task. He is proposing nothing more and nothing less than a change in the foundations of the most prolific physical theory of the last century. The advantage that Penrose has is that he is one of the greatest experts on quantum mechanics, so he knows very well the magnitude of what he is proposing.

Even taking these aspects into account, it is worth asking, does Penrose intend to give a definitive answer? It is easy to realize that what he really aspires to is to continue posing problems rather than to provide solutions that do not allow discussion (Penrose, 1991: 24).

[60] One of the most controversial and elaborate contributions is the twistor theory created by himself. For a recent study of the philosophical implications of such theory, see Heredia (2023).

Penrose is aware that proposing a reform of quantum mechanics from the debate on Artificial Intelligence entails some risks. But he still decides to enter, with the consequences that it entails.

So, is there a possibility of computing the human mind and consciousness so that we can put it into a machine so that it could think and act like us? Penrose says no and I have to admit that I myself feel inclined to support him, although with some differences that I will try to make clear throughout this work[61].

Apart from talking about the appropriateness of science to support his arguments, Penrose also highlights an aspect that supporters of Artificial Intelligence seem to overlook: science can be a double-edged sword. The enthusiasm for obtaining the expected results causes us to lose sight of the possible consequences of these results. Is it so impossible that the situation becomes increasingly complex with the advancement of technology? Penrose is clear that not and that is why he thinks it is strictly necessary to be responsible. But if machines were to evolve to the point of being independent, who would we ask for this responsibility? Penrose understands that the responsibility for everything that may happen must fall on human beings. Possible advances depend on human consciousness and not on that of computers (they will never have it!).

In a debate it is necessary to know the other points of view in order to be able to argue and counterargue based on what each of them defends and rejects.

4. ANOTHER ASPECTS OF THE DISCUSSION

4.1. DIFFERENT POINTS OF VIEW THAT PENROSE TAKES INTO ACCOUNT

Penrose himself realizes that it is necessary to make a classification between the different points of view of the debate so as not to disperse in his criticisms. This is the classification he offers in SOTM:

[61] The ideas developed, as seen in some of the previous quotes, follow the works of Penrose that have transcended the most in the field of philosophy. These are *The Emperor's New Mind* (1989) (henceforth ENM) and *The shadows of the Mind* (1994) (henceforth SOTM).

> A. All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations.
>
> B. Awareness is a feature of the brain's physical action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness.
>
> C. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally.
>
> D. Awareness cannot be explained by physical, computational, or any other scientific terms (Penrose, 2012: 26)[62].

We have that A belongs to the so-called strong (or hard) AI. This point of view defends that the mysteries that the human mind and consciousness hold [without exception] are capable of being known. That is, we can have full power of both the mind and consciousness, to the point of being able to introduce them into a machine and have it acquire these capabilities. At first it may seem that this doctrine advocates a materialist explanation of consciousness and the mind. It would only be necessary to find the appropriate devices to carry out the intended work. However, Penrose correctly points out that in A he intercedes in favor of the role of information[63] rather than matter (Penrose, 2012: 28).

The material can become a secondary element, a mere "pattern of information" that simply responds to what its mathematical programming dictates. Model A will be the one upon which Penrose directs most of his criticism. In any case, this does not prevent him from recognizing that he feels admiration for it, since the main goal of this doctrine is to

---

[62] I consider it necessary to clarify that within the different points of view that Penrose offers us there are conflicts, to the point that currents that in this classification fall into the same block of thought are considered disparate. I think it is necessary to take this into account, but it is more practical (Penrose probably also did it for this reason) not to begin an exercise of division and subdivision between points of view according to what type of nuances differentiate one from the others, since this It would be counterproductive to the purpose of the exposition intended here.

[63] Among the different meanings that the concept "information" has, here we must understand it as the series of knowledge that is introduced into the machine so that it constitutes the structure of its consciousness. As we can see, this topic is not without controversy, since concepts are used that involve many clarifications. For this reason, I think it is best to understand it in its most general form. For actual and concrete works about information and desinformation see Palomo (2021).

get all the possible profit out of scientific work. However, it can also happen (as in fact happens) that it is at the opposite end in the other aspects of this particular point of view.

Model B corresponds to the commonly known weak AI. In B it is argued that the behavior of physical objects can respond to a computational operation, just as it happens in A. The difference between these two points of view is that for A, when a machine responds to its computer program, it is doing so "consciously", while in B it is not, or at least it is not clear that it is. The main reason for this nuance is that the physical (that is, material) composition of the brain cannot be extrapolated to what the machine possesses[64]. Therefore, we can see that the secondary role that materiality had in A becomes a main one in B, causing the most notable difference between both points of view. On the other hand, although it belongs to the same aspect, we also have that computing, essential in A, now takes on a rather futile role in B (Penrose, 2012: 29). Computing serves to simulate consciousness and matter is what allows us to have consciousness.

Point of view C does not have an identifying name in the way that the previous two points of view do. C is the name that Penrose gives to his own view and that is how it is known.

Finally, we have D. In D we find a position that rejects any response coming from the scientific field, a condition that makes it related to mysticism (Penrose, 2012: 26). By arguing that it is impossible for science to solve the mysteries of the human mind and consciousness, the best thing it can do is remain silent. Although in principle this vision clashes head on with the model proposed by Penrose (in which the role of science is crucial), this does not prevent both positions from finding a common point. In C there is a disenchantment with current science. So much so that this leads him to deny the results obtained by said science. Is there the same distrust regarding science? Well, we would have to respond with a certain trick: yes and no. Yes, in the sense that it is a fact that science currently does not allow for conclusive answers

[64] Let's understand that this machine behaves as a human being would.

and this keeps it impotent in this particular debate. And no, in that Penrose does not plan to abandon science, but believes that it absolutely needs a remodeling (Penrose, 2012: 30).

Definitely, D distances itself from both C and A in this section, but not so much from B. Model D is the only one that allows it to adapt to everyone. It is obvious why: this point of view is the least daring.

Let's continue seeing common and different aspects between these models.

4.2. SOME MORE POINTS OF VIEWS

Penrose has been criticized for the fact that in a debate of such magnitude he inadequately contemplates the different points of view involved in this debate. One of the most notable criticisms is that carried out by the philosopher Aaron Sloman in his article "The Emperor's Real Mind"[65] (in clear reference to Penrose's *The Emperor's New Mind*).

This work contains two main criticisms: i) Penrose's lack of precision when talking about the supporters of Artificial Intelligence, ii) the rethinking of the Penrosean use of Gödel's theorem. In this work we will only see the first criticism, since it is the one that is related to what we have been seeing[66].

Sloman considers that the way in which Penrose refers to AI supporters is clearly deficient, since it includes different points of view as one, without taking into account the nuances that differentiate them from each other (Sloman, 2018: 4). Those who defend Artificial Intelligence, therefore, cannot all be limited to the same group. For this reason, Sloman offers a new and more extensive classification of the different ways of defending this point of view, all of them included within strong AI. This classification consists of nine different theories.

[65] Although the original article is from 1992 and was published in Artificial Intelligence, I have used the electronic version with a list of contents and a new post scriptum published in 2018, which is available at the following URL: https://www.cs. bham.ac.uk/research/projects/cogaff/sloman-penrose-aij-review.html

[66] Regarding current studies on the topic of the second critique, see Berto (2009) or Heredia (2019).

Some of them, he argues, are treated by Penrose, although also inadequately. Those developed by Penrose are the most extreme and Sloman calls them T1 and T1a. These two theories argue that the key to everything lies in a single algorithm that has not yet been discovered. This algorithm would be very subtle and also capable of being known. Sloman believes that treating this type of theory seriously is a useless task, since the defense of such a principle is absurd. And the rejection is not based on a personal opinion (that too) but because practically no one who defends the possibility of an artificial intelligence holds such ideas on this particular principle.

Then there is T2. This theory is related to enactivism. Due to the little development that this doctrine had at that time and because it does not make it clear that everything depends on an algorithm, Sloman believes that it is too early to subject it to an in-depth analysis, since it has not yet been developed, so it is not possible to be discussed properly. Nevertheless, Sloman suggests that this theory has much more potential than the previous two, despite the fact that he contemplated at some point the importance of an algorithm with a specific singularity.

This all changes with T3. This theory, although it does not deny the existence of such an unknown and "special" algorithm, differs from the others by defending that apart from such an algorithm there are also multiple interacting computational processes. That is, it does not rule out the composition of various algorithms.

The multiple composition of T3 algorithms makes one aspect clear: simple systems (with a single algorithm) give way to complex systems. T4 is another example of this. This theory establishes that mental states are produced from collections of computational processes carried out on distributed collections of processors[67]. It is not only important to understand that mental states are caused by complex algorithmic systems, but also that in the environment in which they occur they also gain notoriety. This does not mean to refer to materiality, but to the form of composition that allows the circulation of collections of

[67] Let's understand "processors" as electronic circuits.

computational processes. For its part, T5 is very similar to T4, with the particularity that it would allow the design of an intelligent agent to require elements that are not necessarily computational. According to Sloman, this last theory is so vague by not stating what type of non-computational elements would be necessary that it runs the risk of not being really interesting as a theory to take into account (Sloman, 2018: 19). In any case, it is not declared as such.

T6 brings issues hitherto not covered by previous theories. The most important of them is the relationship that mental states can have with the environment. This theory handles the possibility of simulating the physical world on a computer. The purpose of this is to see if a mind created from computational processes that, they claim, are like humans, would act in the same way as humans. The problem arises when trying to simulate the physical environment in computational terms, since it may be the case that there are features in the physical world that escape computable processes. In the case of encountering such non-computable features, the simulation would not be possible. T6 assumes that in the interaction of human beings with the environment there are no essentially continuous processes.

T7 establishes that it is possible to introduce sets of mental processes into a computer as long as it is done in shared time. Time sharing is a process that allows multiple users to run multiple programs on a single computer at the same time, facilitating the breadth of the field of activity of computational processes

T8 is more radical in denying that a single algorithm is responsible for mental processes. These are too complex for their activity to depend on the work of an algorithm, no matter how subtle it may be. Instead, it suggests that this task could be entrusted to the implementation on a network of computers.

The way these theories of Artificial Intelligence try to approach mental processes is to add elements that contribute to their complexity and, therefore, their similarity. But do they really contribute anything new that invalidates Penrose's arguments? Penrose makes the mistake of identifying all AI supporters with those who seek a single algorithm,

ignoring AI views that rule out such a search. But this mismatch will only belong to ENM. In fact, in order to alleviate such an error, in SOTM (Penrose, 2012: 27) he makes explicit mention of Sloman's article.

In SOTM, Penrose continues to consider all supporters of the different types of strong AI from the same point of view (A). Is Penrose still making the same mistake? Personally I think he is not. We can start from the idea that Sloman's classification needs to be taken into account. But it is still true that the nine theories share a final answer in computational terms, whether they are simpler or more complex. It is also worth noting that more complex theories are less consistent, so conducting a separate debate with them is meaningless.

What is important is what Penrose wants to argue about the possibility of the computability of consciousness and the human mind. And therefore in the following section we will see an example that makes his position very clear.

5. CONCLUSIONS: MACHINES AND "UNDERSTANDING"

Turing argued that machines can imitate parts of the human being and that it was a matter of time before this imitation remained a simple anecdote because machines would be able to think like humans do. One of the first aspects in which he was interested in advancing the thinking of machines was the skill that they could acquire in board games. Since then, very good results have been obtained, such as the example of chess modules. Chess occupies a privileged place, since it is a game in which the use of intelligence is essential.

The beginnings of chess modules were not easy, although in a relatively short time they did not struggle anymore. Machines would not only manage to defeat any human, but some would achieve victories over great champions, to the point that today not even the highest level Grandmasters can cope with the best developed engines.

But are these types of achievements conclusive enough to determine that expert chess machines are intelligent and think? Penrose will

answer negatively and to account for this he presents an example in which an expert chess machine shows that it does not understand[68] the game itself beyond what it was programmed to do.

Such an example involves the case of a powerful chess expert computer, known as Deep Thought.

Once this machine had demonstrated its abilities to play chess at the highest level, its understanding of the game had to be tested. The results, however, were more striking than expected. Penrose highlights the result of a test in a particular game. This was arranged as follows:

The machine, which handled the white pieces, had a barrier of pawns arranged on the board in such a way that it prevented the passage of the black pieces (which were also placed in the form of a barrier with their pawns, thus preventing the movement from other figures such as its bishop and its two rooks) towards the white king (which was the only figure it had apart from the pawns). This barrier was formed in such a way that one of the white pawns had the opportunity to capture a black rook. This movement would cause the barrier to break and, consequently, the only defense that its king would have. The "smartest" move in this case for whites is to continue moving the king until the game had to end in a draw.

What was the movement that the computer made? It chose to capture the rook, thus sacrificing the entire game. The result, however, should not be surprising. Deep Thought was programmed to attack when it had the chance, and indeed, that is what it did. And this is precisely Penrose's argument. The machine cannot go beyond its program. The intelligence that it displays is just the result of abilities that were introduced to it from the beginning. But is it right to condemn Deep Thought's possible intelligence for making this particular mistake? Are humans exempt from making mistakes? It is evident that humans make mistakes, and many, but there is a feature in Deep Thought's error that should not be overlooked. Many humans could have made the same mistake as the

[68] Let's understand this concept as we saw it in note 5. That is, in relation to the terms knowledge, intelligence and consciousness.

machine, but we would not be talking about the same case. Deep Thought is a chess expert, so the equivalent situation regarding humans would be knowing how many chess-expert humans would make the same mistake. The answer is none, because they understand the game of chess and the move was clear enough to direct it towards a draw!

This last example should not be misunderstood. Today no module would carry out the move that Deep Thought did. But the point is not the mistake itself, but the difference between how humans think and how machines do. We can see this with an example in which the machine, instead of making a mistake, makes a series of brilliant moves. If we take an engine from today that finds a checkmate thirty moves in advance, of course denying it any kind of intelligence is not correct. However, it is pertinent to say that their intelligence is different from that of a human being (because no one foresees a checkmate with such advance notice!). This is Penrose's defense and one with which I completely agree.

It is obvious that Penrose has his convictions, and it is easy to see that he does not try to give absolute answers, since this seems like a futile task. Therefore, it is best to sit down and debate to find what brings us closer and what distances us.

6. ACKNOWLEDGEMENTS

The author is supported by NextGenerationEU founds, with a contract within “Margarita Salas” program.

7. REFERENCES

Álvarez, P., Bañares, J., Latorre, P., Velilla, S. (2005). Programación, Zaragoza, C.P.S. Universidad de Zaragoza.

Berto, F. (2009). There´s something about Gödel: The complete guide to the Incompleteness Theorem, Oxford, Wiley-Blackwell.

Copeland, J. (2017). “Turing´s great invention: the universal computing machine” in Copeland, J., Bowen, J., Sprevak, M., & Wilson, R., The Turing Guide, Oxford, Oxford University Press, pp. 49- 56.

Floyd, J. (2017). “Turing on “Common sense”: Cambridge Resonances” in Floyd, J. & Bokulich, A., Philosophical Explorations of the Legacy of Alan Turing: Turing 100, Boston, Springer, pp. 103-152.

Gardner, H. (2006). Multiple Intelligences: New Horizons, Nueva York, Basic Books.

Heredia, D. (2019). “La importancia del teorema gödeliano en el pensamiento de Roger Penrose”, in Naturaleza y Libertad. Revista de estudios interdisciplinares, (12), pp. 159-178.

(2023). “On the possibility of a real reform in current physics: Penrose and twistor theory”, in Perspectivas, 7(2), pp. 49-71.

Palomo, M. (2021). “How disinformation kills: philosophical challenges in the post-Covid society”, in History and Philosophy of the Life Sciences, 43, (51), pp. 43-51.

Penrose, R. (1991). La nueva mente del emperador, trans. by J. García Sanz, Barcelona, Grijalbo Mondadori.

(2012). Las sombras de la mente, trans. by J. García Sanz, Barcelona, Crítica.

Poole, D., Mackworth, A., Goebel, R. (1998). Computational Intelligence: A logicial approach, Oxford, Oxford University Press.

Russell, S. J., Norvig, P. (2004). Inteligencia Artificial: Un enfoque moderno, trans. by Corchado Rodríguez, J., Martín Rubio, F., Cadenas Figueredo, J., Hernández Molinero, L., Paniagua Arís, E., Fuentetaja Pinzán, R., Robledo de los Santos, M., y Rizo Aldeguer, R., Madrid, Pearson Educación S.A.

Sloman, A. (1992). “The Emperor's Real Mind”, Artificial Intelligence, 56, pp. 355-396.