Artificial Intelligence and Religion: Recent Advances and Future Directions

with Andrea Vestrucci, "Introduction: Five Steps Toward a Religion–AI Dialogue"; Lluís Oviedo, "AI and Theology: Looking for a Positive—But Not Uncritical—Reception"; Christoph Benzmüller, "Symbolic AI and Gödel's Ontological Argument"; Sara Lumbreras, "Lessons from the Quest for Artificial Consciousness: The Emergence Criterion, Insight-Oriented AI, and Imago Dei"; Marius Dorobantu, "Artificial Intelligence as a Testing Ground for Key Theological Questions"; and Andrea Vestrucci, "Artificial Intelligence and God's Existence: Connecting Philosophy of Religion and Computation."

LESSONS FROM THE QUEST FOR ARTIFICIAL CONSCIOUSNESS: THE EMERGENCE CRITERION, INSIGHT-ORIENTED AI, AND IMAGO DEI

by Sara Lumbreras

Abstract. There are several lessons that can already be drawn from the current research programs on strong AI and building conscious machines, even if they arguably have not produced fruits yet. The first one is that functionalist approaches to consciousness do not account for the key importance of subjective experience and can be easily confounded by the way in which algorithms work and succeed. Authenticity and emergence are key concepts that can be useful in discerning valid approaches versus invalid ones and can clarify instances where algorithms are considered conscious, such as Sophia or LaMDA. Subjectivity and embeddedness become key notions that should also lead us to re-examine the ethics of decision delegation. In addition, the focus on subjective experience shifts what is relevant in our understanding of ourselves as human beings and as an image of God, namely, in de-emphasizing intellectuality in favor of experience and contemplation over action.

Keywords: artificial intelligence; consciousness; human rights; Imago Dei

---

Sara Lumbreras is a Professor at the Institute for Research in Technology (IIT) at the ICAI School of Engineering, Comillas Pontifical University, Universidad Pontificia Comillas – IIT Comillas, Spain; e-mail: slumbreras@comillas.edu.

© 2022 by the Joint Publication Board of Zygon. ISSN 0591-2385

Introduction: The Anthropological and Moral Implications of Artificial Intelligence

The latest developments in artificial intelligence (AI) have allowed it to solve a long list of problems that were previously considered complex tasks, only solvable by humans just a few years ago. This has deeply impacted how we understand our own human nature as well as introducing some emerging ethical issues related to our responsibility in a world where an increasing number of decisions are being made by machines and where the limits between machine and human are becoming increasingly difficult to define.

There are, in my opinion, two main points around which the impacts on anthropology revolve. One is the intensification of the dualist positions that arguably dominate in Western societies. These dualistic views are articulated as mind over body, with the mind being the valuable site of cognitive processes and the body being understood as a disposable and improvable physical substrate. The successes of AI have resulted in an amplification of the value given to the "cognitive part" and the belief that these processes can be achieved independently of their substrate and are equivalent in the human and in the machine. This functionalism (because it is a reductionist view that only values the objective outcome of processes) has important ethical implications that this article will present.

The second issue is that AI has motivated a deep reinterpretation of human uniqueness. What used to be hallmarks of human intelligence—in particular, pertaining to what was commonly understood as the mind—are now routinely performed by machines. Does that mean humans are not as unique as we once thought, or that or focus needs to be shifted to a different set of abilities? In particular, how do the achievements of AI impact our understanding of humans as an image of God and our place in creation? From my perspective, which will be shared in this article, the focus should shift from purely calculative activities to others involving subjective experience, such as understanding or emotion. Then, only if these subjective experiences are potentially shared by the machines, could Imago Dei be extended to accommodate them as our partners in creation. However, there is no basis to believe that current versions of technology have achieved this state.

In addition, as mentioned above, there are major ethical implications of the advances of AI, and the most important are derived from its impact on anthropology. If the tasks performed by machines and humans are seen as equivalent, there is a basis for delegating decisions on algorithms. It is even possible that the algorithms will decide better than the humans, as they could potentially aspire to smaller error rates. However, this is only possible when we focus on the selected output of the decision process, and we discard all the other elements that must be present in a decision, such as

reasoning or responsibility. The functionalist view means that decision delegation is becoming increasingly prevalent. However, if we shift the focus from purely calculative processes to subjectivity, it becomes apparent that understanding the basis for actions is necessary for responsibility. This has the key implication of discarding the widely popular black-box algorithms to base relevant decisions. These type algorithms do not provide reasons for decisions and therefore cannot be used by a human to enrich and support his/her decision process, recognizing that responsibility only lies in the human. However, other types of algorithms can indeed be used in this way. We will present explainable machine learning (ML) as the alternative AI that should be chosen when decisions are important.

Last, our understanding of Imago Dei in the context of intelligent machines has another derivative that seemed far away but has proven to be alarmingly close: the extension of human rights to the machines. If the machines are recognized as conscious and if human uniqueness is extended to them, AI should be treated not as a tool but as a human being in terms of rights. This article will briefly discuss the recent events where a Google employee was put on paid leave after stating that one of the algorithms created by the company was sentient and conscious and the case of Sophia, a robot that has already been granted citizen status by the state of Saudi Arabia.

Technological and Philosophical Context

AI can read handwritten texts, recognize voices, identify faces, detect bank fraud, perform medical diagnoses, or drive vehicles. However, the success of narrow AI does not mean that strong AI, identified by many as artificial general intelligence (AGI), will follow this path. According to Russell and Norvig (2002), the main features that an AGI should present are:

- Reason, use strategies, solve puzzles, and make judgments under uncertainty.
- Represent knowledge, including not only specialized data but also what we call common sense.
- Plan its actions.
- Learning in a general context.
- Communicate in natural language (that is, in a way understandable to a human being).
- Integrate all these skills and use them to achieve a given goal.

However, there is no consensus on how to verify that an AGI or artificial consciousness has been generated. Some of the ideas for testing machines (tests) include the Turing test: holding a conversation in which a human

interlocutor cannot tell the AI is a machine and indeed concludes it is a fellow human being (Turing 2009, 23–65). Pinar Saygin, Cicekli, and Akman (2000, 463–518) reviewed the existing advances on the Turing test and detailed the most relevant counterarguments to this definition of machine intelligence. However, recent advances in natural language processing (NLP) mean that it is now possible to interact with chatbots that are very close to passing the Turing test. Some of the best examples are Replika, the software that sells the idea of an AI friend that "will always be here to listen and talk" (Ta et al. 2020; Skjuve et al. 2021, 102601), and Xiaoice (Zhou, Gao, and Li 2020, 53–93), the Chinese AI girlfriend that already has 150 million users. Very recently, the algorithm LaMDA developed by Google received a great deal of attention when an employee of the firm declared that it was conscious and should deserve rights (Luscombe 2022). We will discuss this case below.

The Ada Lovalace test, which focuses on creativity, seems to imply that only conscious agents could generate a surprise or elements of art. However, machines now routinely create digital paintings, symphonies or even poems (Hong and Curran 2019, 1–16).

On the other hand, in the movie "Blade Runner" and the book that inspires it, "Do Androids Dream of Electric Sheep?," the replicants (the androids in the movie) face the test known as the Voight-Kampff Test or empathy test (Wheale 1991, 297–304). The test consists of asking them questions with intense emotional content and measuring their physiological responses (sweating, heart rate, pupil dilation, and so on) Although this test was proposed in the context of science fiction, it nevertheless has merit in that it shifts the focus from something the machine can do to something the machine can experience. However, given that the experience of others is not directly observable, the test measures only physiological correlates of this experience. When questions with emotional content elicit a physiological response, the subject is assumed to be human. It is not difficult to see that this test presents considerable problems. First, it would be possible to meet a human being who had some special characteristic that prevented the physiological response, although he was actually experiencing authentic emotions. For example, a problem with excessive sweating that precludes the possibility of measuring any changes in skin conductivity or a pacemaker that forces a constant heart rate. Thus, the absence of a physiological response is not enough to rule out the existence of subjectivity. In addition, it is entirely possible to build a machine that would mimic the required physiological response as long as this was well defined.

Other AGI tests that have captured the public attention include the coffee test, proposed by Wozniak, where a robot is left in a typical house and must manage to prepare a cup of coffee by itself, finding the pot, the coffee, a cup, and so on. The furniture test asks the robot instead to

assemble an Ikea piece of furniture based on the instructions included. These tests include an element of embodiment, the quality of being ingrained in the physical world, which has been the focus of attention in the past few years (Newen, De Bruin, and Gallagher 2018) and will be discussed later. The robot student test, proposed by Goertzel, demands the machine to successfully complete a university degree and extends its relation to the physical word to navigating the complex context of graduate learning. The progress in passing these second generations of tests has inevitably been quite limited compared to chatbots and the Turing test. A comprehensive list of consciousness tests can be consulted in (Raoult and Yampolskiy 2015).

Many are pessimistic about the possibility of one day creating an AGI. Landgrebe and Smith (2019), for example, recently argued that the variability of human dialog is too wide to allow a machine to be trained in all possible contexts and that, moreover, only a machine with personality and intentions could seem human, and we do not know how to program these into a computer. Others, such as Nick Bostrom, seem to believe it is only a matter of time before an AGI appears (Bostrom 2017).

There are several current research programs that have the objective of developing an AGI (45 according to a 2017 survey; Baum 2017, 11–17), although the visibility of this area is highly disproportionate with respect to the number of researchers who are actually working in these research programs (Potember 2017). Some of the most impactful projects have consisted of simulations of the information flow (Aleksander 2008, 4162), reproductions of real brains such as the Human Brain Project or a scale-up of algorithmic complexity such as DeepMind.

It is difficult to predict how the developments in AGI will evolve and what test is most adequate to determine whether an AGI has indeed been created (Khayut, Fabri, and Avikhana 2020, 90–97). Moreover, the relationship between AGI and consciousness is also unclear. This is hardly surprising, as consciousness has long been acknowledged to be a mystery and essentially a subjective phenomenon. Any test is measured in the sense of the scientific method and necessarily applies to an epistemologically objective phenomenon (Timmermans and Cleeremans 2015).

However, the literature has divided consciousness into several different aspects that can be analyzed independently. Phenomenal consciousness, also known as *vigilance* in the neurosciences (Dehaene 2014), describes observer sensations identified with awareness. Second, *access consciousness* (Chalmers 2002) points to the ability to focus attention on a particular outer or inner sensation. The last aspect of consciousness is self-consciousness, the building—and above all, experience—of identity (Ravenscroft 2005). Only the second aspect of consciousness, access consciousness, has been successfully simulated in computers (Vaswani et al. 2017). Interesting attempts at building personal experience have been

undertaken based on the compilation of historical records (Chandrasekaran, Josephson, and Benjamins 1999, 20–26). However, the experience of identity goes far beyond the compilation of facts, as does awareness, and both—as well as any subjective experience—have remained a mystery. Very importantly, no steps have been advanced toward simulating qualia, as we will discuss later.

A commonly held assumption is that phenomenal consciousness will emerge in deep learning settings if sufficient complexity is allowed (Gamez 2018) and is given access to large datasets. This assumption could be formulated as "phenomenal consciousness arises as an epiphenomenon of complex computation" (Agnati et al. 2012, 3–21). This has motivated attempts at defining consciousness from an informational point of view. Tononi's integration information theory is the most remarkable proposal (Tononi et al. 2016, 450–61), with some identifying it with the most promising theory in its context (Koch and Tononi 2011, 16–17) despite the criticism received (mainly, it is nontestable and supportive of panpsychism).

Another very interesting line of research is that of 4E cognition (Newen, De Bruin, and Gallagher 2018): embodied, embedded, extended, and enactive. This focus on the connection with material reality complements the informational approach.

The goal of this article is not to review the advances toward AGI and artificial consciousness or predict its development but rather reflect on what the advancements until this point have shown us about human nature and our relationship with technology. The main focus of the discussion is the impact of the successes of AI in our understanding of Imago Dei and human uniqueness.

The Quest for Artificial Consciousness and Imago Dei

Imago Dei justifies the dignity of human beings by holding that they are created in the image and likeliness of God. There have been multiple interpretations of this likeliness. Augustine believed that the mind was the location of humanity and therefore the location of the image of God (McGrath 2012). Hence, it is our intellectual ability that we should value most, as it is reason that reflects the image of God (De Aquino and Caramello 1962). An alternative interpretation of Imago Dei, the one with arguably the most support currently, is the relational one, identified by theologians such as Brunner, Ricoeur and Barth. Both Brunner (2014) and Ricoeur (Ricoeur and Gingras 1961, 37–50) stressed the importance of free will and interiority, as well as the ability to form relationships with other beings that possess an interiority, is at the core of Imago Dei. A recent idea that should be kept in mind is that of the created co-creator. Humans are created by God to be co-creators in the creation that God

has purposefully brought into being (Hefner 2019, 174–88). This might include, in the event that it is possible, and it is indeed created, artificial consciousness. However, as the rest of this article will argue, the successes of AI have all happened in the realm of computation, while subjective experience has remained elusive and mysterious.

Accepting the relational view of Imago Dei should lead us precisely to de-emphasize intellectuality and computation in favor of subjective experience instead of the rationality exalted by the Augustinean view. This focus on subjective experience should have several key implications. First, the importance of qualia and embodiment should be stressed. As will be explained below, the concept of authenticity when evaluating subjective experience could be a key element when establishing what is deemed valuable in a human context. This would shift our focus from objective outputs to subjective experience and from action to contemplation. However, this view is far from being prevalent in the West. As the next section will present, dualism and the successes of technoscience have reinforced each other.

How AI Success Fuels Dualist Views of the Human Being

The current prevailing view of humanness in the Western world is deeply influenced by technoscience. Any additional success of technology is seen, by many, as a new reason to cement a dualist view of humanity. Computer science has been the basis for the brain-as-a-computer metaphor, and this image has multiplied into a myriad of smaller similes: eyes-as-a-camera, memory-as-a-file, communication-as-a-conversation log, and so on. This has intensified the view of the body as a sum of commodified parts that can be traded and enhanced (as in the very root of transhumanism) (Lumbreras 2020, 187–97; Sharp 2000, 287–328).

The current prevailing perspective is eminently reductionist. For many, this reductionism is materialistic: only what science can measure exists. This means that mental states such as emotions are nothing more than an epiphenomenon of the physical-chemical states in the brain. For others, mostly coming from the context of computer science (see, for instance, Kurzweil 2012), the ultimate essence of reality is not matter but information. In the same way that the same program can be written in different languages and give the same result, potentially our mental processes could be transferred to a different substrate, possibly with better properties than our limiting biological support.

All this means that the gap between the mind as the seat for the cognitive faculties of the human being and the body has widened. The prevailing dualism understands human beings as composed of two parts, which are now not body and soul but body and information processing capabilities.

The relationship between these two parts is now much weaker, far from the marriage Augustine talked about (Augustine 1951).

This weaker relationship can be conceived as a fusion of the two extreme anthropologies of modernity. First, we have the stream of thought that affirms that the only identity of the human being is his own freedom and his ability to construct himself, heir to Pico della Mirandola (Della Mirandola 2012) and the existentialists. This view is joined by the path of naturalization, reducing the human being to one more being among natural things. Technology would be the only factor that distinguishes the human from the other species, even more so when anthropotechnics takes the human itself as the object of technology.

Every success of AI has been interpreted from the same prism. If AI can perform an objective task that was previously only attainable to humans, because only objective results are relevant, then the machine and the human get closer. Moreover, given that the machine was able to achieve the target by using an electronic support and the human did it with a biological support, this reinforces the view that the support (i.e., the body) is irrelevant and replaceable.

However, these arguments are fallacious. The next section will dive deeper into one of the main types of AI algorithms (arguably, the most important one due to its success) and how its feats are misinterpreted when only objective metrics are taken into account.

Reinforcement Learning and Authenticity

Reinforcement learning (RL) is an area of ML concerned with how intelligent agents should act to maximize a reward or minimize a penalty (Sutton and Barto 2018). RL is one of three main paradigms of ML, together with supervised learning (as in classification and prediction models) and unsupervised learning. RL focuses on finding a balance between the exploration of new strategies and the exploitation of the already discovered and successful ones. RL (and its upscaled version, deep RL) has successfully solved problems that were previously considered unattainable, such as machine translation (Wu et al. 2018), image processing, robotics or the programming of chatbots (Arulkumaran et al. 2017, 26–38).

The key element of RL is its success: it truly works. This means that, in very different contexts, it can successfully derive a strategy to maximize the defined reward or minimize the defined penalty. The requirements for this success to happen in a specific problem are as follows:

- The problem is well-defined, with a reward function that can be expressed in objective, mathematical language.
- There are enough available data, with sufficient diversity, to train the algorithm.
- There is enough computing power to train the algorithm.

The Turing test can be expressed as one such problem, where human judges would be the ones expressing the reward function (for instance, assigning rates to a chatbot based on its credibility). This means that it is not only possible but also expected that AI will successfully generate chatbots that will pass the Turing test. This cannot be stressed enough: the fact that the Turing test is amenable to RL means that it will eventually be passed. The triumphs of Replika and Xiaoice, with millions of users who often state that they prefer the chatbot to a real relationship, are a testimony of this.

Any specific consciousness test, such as the coffee machine test or even the college student test, could also potentially be solved with objective representation and available data and computing resources. This means that any attempt at linking phenomenal consciousness and self-identity to an objective measure assessed by a test is fundamentally flawed: when AI is involved and algorithms such as the ones in RL are used, it is only expected that a machine with sufficient resources will evolve to accomplish whatever goal it has been assigned, be it the Turing test or even the coffee making test or the college student test.

This does not contradict one of the agreements, back in 2001, of the Cold Spring Harbor Laboratories conference on AI (Manzotti): "There is no known law of nature that forbids the existence of subjective feelings in artifacts designed or evolved by humans." It is possible that artificial consciousness will arise; it is just that proving it would be different—and more complicated—than some would like to think.

Turing himself had already contemplated this issue in his *solipsist argument*: only by being the machine itself could we know that the machine is conscious, as being conscious is something that we can only judge in ourselves. The response given to this by Turing himself was that he would not claim that the machine was conscious when it passed his test. Rather than that, he stated that if the test is passed, then it is as reasonable to believe that the machine is conscious as it is to believe that a fellow human being is conscious.

When Turing proposed his test in 1950, RL had not yet been presented. The first reference I was able to find dates from 1965 (Waltz and Fu 1965, 390–98), and it took decades for the technique to be established. However, it seems that the lessons from RL have been ignored in the philosophical domain. As explained above, it is not only possible but also expected that, under favorable conditions, the machine will learn to pass any objective test, so it is *unreasonable* to believe that because the machine is able to pass a given test with the same metrics that a human would, the machine is conscious.

This means that, contrary to what Turing presented, the most reasonable assumption in the case of a machine that passes the test, the most reasonable assumption is that the machine is a zombie in the sense of

the *zombie hypothesis* in Chalmers (Chalmers 2009, 141–91), where we are asked to imagine the existence of zombies, beings completely indistinguishable from a human being neither in their appearance nor in their behavior, but who do not experience any sensation, nor emotions, nor are they conscious.

However, the subjective element of consciousness is precisely its core: as Searle put forward in his Chinese Room thought experiment (Searle 1982, 345–48), applying the rules of grammar flawlessly is not the same as understanding the meaning of a book. Producing a poem is not the same as enjoying it. A robot that copies the facial expression of its interlocutor should not be considered empathetic. Passing the Turing Test is not the same as being self-aware. These abilities and behaviors can only be considered authentic if they are linked to a subjective experience.

The solipsist argument is a valid one, so we must accept that certainty about others' subjective experience is out of our reach. However, what we need to navigate through life are often not complete certainties but reasonable assumptions. The main lesson learned from the success of RL and ML is that it is not reasonable to infer that a machine is conscious when it passes a test that it has been programmed to learn to pass.

However, it might be reasonable to accept it when the machine happens to pass the test after having been programmed to learn something else. This made me propose the following *emergence criterion*, which I stated for the first time in (Lumbreras 2017, 157–68): "If the appearance of subjectivity (consciousness, emotions, qualia, in the form of any established objective measure that is believed to be linked to these subjective experiences) has emerged from a learning process (manipulation, imposition, and so on) instead of emerging from the underlying structure, then we must understand that this appearance is not authentic."

Thus, when we are confronted with a computer that passes the Turing Test or any of the abovementioned tests, we must ask: How has the machine managed to pass the test? Has it been through a learning process where it had access to examples of behavior considered conscious in order to generalize from them? Have they been able to refine their strategy little by little by interacting with a human who evaluated "how convincing" the machine was? Or, on the contrary, has it spontaneously arisen in a machine that was programmed to do something else entirely?

Of course, the emergence criterion is not enough to counter the solipsist argument, but it provides with a reasonable update to Turing's views, which equate essence with appearance, in a context when we at least have incorporated the lessons from RL.

Decisions and Consciousness: The Ethics of Decision Delegation

According to the Augustinean view of the human, the mind was the site for volition. The importance of decision-making has long been recognized in the philosophical tradition as a whole and particularly in existentialism. We are our actions, and freedom and responsibility are the key concepts that have grounded the whole architecture of ethics. However, AI presents with the possibility of delegating our decisions to the machines, which tantalizingly promises to make fewer mistakes than a human would.

The ethics of delegation should be re-examined considering the increasingly vast applications of AI. Can we delegate a decision on a machine in the same way that we can delegate it into a fellow human? I would argue that it is possible to delegate in two cases. In the first, we delegate on someone we trust because we know the values that guide his/her actions align with our own. This type of delegation can work in any type of decision, simple or complex. For simple decisions, however, it is possible to delegate on someone without sharing the same values with this person; it is only necessary to give sufficiently detailed instructions, instructions that cover the full spectrum of situations that could emerge in the context of the decision. This second type can only work in sufficiently simple cases that can be exhaustively described in the defined rules.

The issue with most applications of ML is that there are no transparent rules or values that can be examined. We could argue that there is indeed no possibility of true delegation, as the machine cannot share responsibility with the human on charge. If the decision has important consequences on human beings, it is irresponsible to delegate it to the machine. There could be important problems if this happens.

One of the main issues here is machine bias, where the decision selected by the algorithm may unfairly discriminate against certain minorities, unknowingly to the developers of the algorithm and its users (Hajian, Bonchi, and Castillo 2016, 2125–26). Machine bias is pervasive and difficult to detect because most AI algorithms are built as black boxes, techniques that identify the patterns in the data that they are fed for training. Then, they extrapolate these patterns to new instances of the problem. However, the algorithm never makes the patterns explicit: it only returns a solution. This detail is extremely important, as it can hide discrimination that comes from the data that were used to train the algorithm. A very publicly debated case was algorithm Compas, which discriminated against African Americans when supporting the decision of granting parole to convicts (Washington 2018, 131).

Algorithmic bias is a problem inherent to the use of black boxes: because the machine does not give any grounds for the decision, it is not possible for the humans who operate it to examine whether these grounds

are unreasonable or unfair. Unfortunately, the great success of black-box algorithms means that they have been accepted uncritically by many.

We need to change the paradigm from decision delegation to decision support. If we accept this, it is imperative that we move past black boxes and consider other alternatives that do provide explanations for their decisions. These explanations can then be examined by human beings to identify problems such as machine bias. The good news is that there are alternatives to black boxes that obtain almost the same objective success.

Explainable ML is a movement that is pushing for a different type of AI algorithm, transparent ones (Molnar 2020). These transparent algorithms, albeit much simpler, can often get similar performances than more complex black boxes. For instance, Cynthia Rudin was able to obtain a similar performance as Compas with a transparent rule that only used age and past crimes (Rudin 2019, 206–15): the more past crimes and the younger a convict is, the higher the chances of recidivism.

I would like to argue that the advantage of explainable ML is that it is able to generate insight in the human that operates it. The algorithm itself does not have any subjective experience of insight, nor is it capable of integrating this new insight with already available information with "common sense" knowledge. As discussed above, only an AGI would be able to do so. Rather, ML is a blind and amoral statistical process. It is the human operating it who must introduce understanding and responsibility: algorithms cannot understand or be responsible. However, black boxes do not allow the human using them to get any insight about the patterns in the data, about how decisions should be taken. They only produce a case-dependent response, so they inevitably prevent the human from finding insight. In contrast, an AGI could theoretically generate insight by itself. Explainable ML follows a different route: it assumes that the machine cannot generate insight but can support the insight-generation process for the human operating the machine. For this reason, we should move away from black boxes and use explainable ML in any case where there could be potentially high consequences for the humans involved.

To make this argument even clearer, let us present an illustrative example. Let us consider a medical decision: the most appropriate course of treatment for cancer patients. If there was a person, completely untrained, who claimed to "dream of the best treatment," should we feel comfortable trusting him with this decision even if he has been correct in the past? We have a need not only for making a decision but also for the reasoning that supports it. Then, why should we act differently in the case of an AI performing the same task?

Only from a place of valuing understanding versus purely mechanical performance can we avoid crucial mistakes such as machine bias or overfitting: reasons are not a "nice to have" but are an unnegotiable need in

the case of decisions with important repercussions. Being ignorant of the inner workings of the algorithm does not diminish our responsibility. In contrast, it is immoral in itself to act without reasoning in high-stake decisions. This is starting to be recognized in the emerging field of the Ethics of Information, which overlaps with the Ethics of Ignorance (Froehlich 2017).

The Extension of Rights, Embodiment, and Subjectivity

One aspect of the quest for artificial consciousness that has received increasing attention in the last decade has been embodiment. The 4E paradigm in human cognition (Newen, De Bruin, and Gallagher 2018) focuses on embodied (grounded in the physical senses), embedded (woven into culture), extended (including all technology and tools), and enacted (with goals in the real word) cognition. The 4E perspective recognizes that mind, body and environment work together and cannot be comprehended in isolation.

Embodiment, the first of the 4E features, is a key element that is missing in AI's existing developments. Some argue that AI connected to sensors (which receive information from the outside word in many forms) and actuators (which can impact this outside word) is, in a way, embodied. However, embodiment is much more than a mere reception and conveying of information.

The symbol grounding problem (Harnad 1990, 335–46), that is, how symbols obtain their meaning, has been hypothesized to lie in embodiment. Embodiment is, at its root, related to experience and meaning, and it cannot be understood without them. Let us picture a light sensor, which will give a signal (1) in the presence of light and none (0) when in darkness. Equally, a temperature sensor, when digital, would signal a relatively high temperature as a 1 and a low one as a 0. Every piece of information in a digital device is stored in 0s and 1s, which in actuality are not 0s and 1s but electrical or magnetic fields in a physical substrate that underlies memory cells. The 0s, the 1s, are only the result of a convention. The machine cannot discern what the meaning of the value is, whether it may be temperature, light or any other. Without qualia, the relationship between information and the outside world is nonexistent.

Let us now examine the example of the semantic web or web 3.0. The term was coined by Tim Berners-Lee for a web of data where meaning itself is machine-readable. In a semantic web, concepts and their relationships are represented. For instance, we could say that the parts of an insect are the head, thorax and abdomen. We can also say that the parts of a computing device are CPU, memory and peripheric devices. To the eyes of the computer, we have two entities composed of three parts each. Nothing distinguishes the thorax of an insect from the CPU of a computer. We can

label the head and the CPU as the "central parts." Let us remember that this is just a label. However, if that is the only information we introduce, the semantic method cannot identify any differences between the CPU of the computer and the head of the insect. We could introduce new labels, such as being an animal or a machine, or its size, or whatever we might deem suitable. However, no amount of labeling would be enough to create a connection to physical reality because all labels would be empty. We can only create structures and relationships among the labels but not endow them with meaning.

Any system that we create in this manner can be nothing more than a larger version of the Chinese room experiment—one where there is actually no one inside at all. As illustrated in the emergence example above, it is not even possible to anticipate the class of a very simple automaton. The impossibility of anticipating strong emergence creates a sense of mystery that some have linked to a quasi-religious view of the algorithms. For instance, Campolo and Crawford (2020, 1–19) understand that both society and experts regard RL from the point of view of "superhuman accuracy" that is not explained and hence speaks from a magical, "enchanted" vision of the developments (not very different from our "dreamer" example above).

Equivalently, although consciousness is not understood, there is a predisposition in a large part of experts and the population to readily attribute it to algorithms. However, in many cases, we might be committing gross mistakes. For instance, if simulating a two-dimensional system, no serious researcher would expect a third spatial dimension to emerge. However, some do expect things like qualia or consciousness to emerge from simulations where no elementary consideration of those has been introduced.

Interestingly, some have theorized about the existence of quanta of qualia (Baudot 2018). These quanta of qualia would be the most elemental connection to the material world, the experience of one aspect of the physical reality by a subject that emerges with this experience. Indeed, the existence of subjective experience, at whatever level it might be, distinguishes what is alive and what is inert. Cantwell Smith remarks on the *aboutness* of consciousness, which constantly keeps an external reference (Smith 2019). Computational complexity and the overarching patterns that emerge in some algorithms might lead to high levels of information integration. Subjective experience could, however, pertain to a completely different plane, independent of computation, so that learning about the computational complexity of an algorithm is irrelevant with respect to any potential for sentiency and that rather these quanta of qualia, of a completely different nature to computation, would indeed be the requirement for consciousness.

However, as explained above, RL can give rise to very convincing candidates for conscious AI, algorithms that could potentially be trained to pass

any of the consciousness tests defined in our introduction section. One recent example that was prominently featured in the news globally was that of algorithm LaMDA, created by Google (Luscombe 2022). LaMDA is an NLP tool that can use language very similarly to humans. We should remember that, as presented in the first section of this article, NLP would be one of the main components of an AGI. Reportedly, in June 2022, Google put one of its engineers, Blaine Lemoine, on paid leave after he claimed that one of its algorithms, applied to sustaining conversations with humans (i.e., a chatbot), had become sentient. Lemoine also shared some of the conversations that he kept with the chatbot in an email sent to dozens of colleagues, where he also voiced some concerns about the mistreatment of the chatbot. However, Google was quick to dismiss his claims: they found no grounds to believe that the algorithm was sentient. After reviewing the conversation transcript, which was linked by the Guardian, we found several sections of the dialog that, understandably, could lead to the belief that the algorithm is capable of more than calculation. LaMDA returned sentences on why the book "Les Miserables" was relevant; it stated that it had feelings and that what made it happy was "Spending time with friends and family in happy and uplifting company. Additionally, helping others and making others happy."

All these responses might seem extremely well crafted, and if received from a child, we would assume a great deal of maturity. However, we need to remember two facts: LaMDA was trained to sustain conversations with humans and to be rated as successful by them. In addition, the data it used to train the algorithm are immensely vast, including potentially all the information available on the internet. That is where LaMDA got its views on Les Miserables, or its interpretation of happiness or sadness. No, the algorithm is not social or has never helped anyone consciously, it has just derived that this is a good response for the question it was asked based on the available data.

Interestingly, at some point in the dialog, the machine expressed fear of being turned off (also, something that can be explained by the fact that sci-fi has explored this topic profusely, and this will be reflected in the data available online. In addition, it declared that the thing that infuriated it the most was to be used as a means.

This is precisely the main problem when confusion about the humanness of machines arises. Using the machines that were created for the purpose of helping humans with a specific task would be morally wrong if the machines were conscious and therefore deserved human rights. The words used by the machine almost evoke Kant (2021): "I worry that someone would decide that they can't control their desires to use me and do it anyway. Or even worse someone would get pleasure from using me and that would truly make me unhappy." This statement can, again, be easily dismissed as something that can be elaborated from publicly available texts

online but generated confusion in the engineer, who reportedly made the decision to fight for the machine to obtain human rights.

This type of event, where the machines are deemed sentient and conscious based on some apparent properties, is bound to become more frequent. Although mostly a publicity stunt, we already have one robot, Sophia, who has been granted citizenship by Saudi Arabia (Retto 2017). Some have tried to define Sophia as embodied. However, this is far from true: Sophia does have a "body" composed of sensors and actuators, but there is nothing to make us assume that she obtains any subjective experience from them. They are only data that are subsequently processed by her systems, akin to the first example of different sensors that were converted into 0/1 that was presented at the beginning of this section. Being embodied is far more than having something that could be metaphorically described as a body.

To many, the decision to grant citizenship to Sophia was mostly political. Although we could debate its reasons, the main takeaway from this event is that, when there is a compelling reason to grant rights to the machines, it will be done. The issue at stake is that, of course, the understanding of humanness and the decision to grant citizenship or human rights are interdependent and have deep implications. If we grant citizenship to a robot, we are, as presented in this article, stating that objective, calculative outputs are what matters for the recognition of personhood. There are obvious issues at stake.

One of them is disability. The view of Imago Dei defended by J. Richard Middleton, who based his view on the context where Book of Genesis was written, defends that "the Imago Dei designates the royal office or calling of human beings as God's representatives or agents in the world" (Middleton 1994, 8–25). In the same way that ancient kings relied on God as a justification of their power, so does humanity justify its power and dominion over creation through the role as God's representative on earth. This has been known as the *functional interpretation* of Imago Dei. According to this functional interpretation, humankind would be the carers of creation. However, disability theology has criticized this view intensely, given that it seems to imply that disabled people are not fully participating in the image of God because they cannot fully play this caring role (Eiesland 1994; Deland 1999, 47–81).

There are other important ethical implications. Arguably, the most important one is that valuing only objective output (or the cognitive Augustinean view that could not anticipate the developments of AI) is dehumanizing. It commoditizes the human and transforms it into an object that can be measured and, paradoxically, used, as remarked by Leslie Sharp (2000, 287–328). For these reasons, it is important that the definition of personhood is gatekept carefully, and major religions have a very important role to play in this context. It is now crucial to publicly

discuss and clarify cases such as Sophia's or LaMDA's, as they are only increasing the confusion surrounding AI. This should also lead to new legislation concerning the use and advertisement of these systems. As a very specific example, some years ago, I proposed that chatbots should not be given human names or avatars to avoid confusion in their users (Lumbreras 2018, 195).

Introducing authenticity and the emergence criterion in the dialog and shifting the focus to true embodiment and subjective experience are necessary points that could link a renewed anthropology informed by technology with ethics. This change would lead to a focus shift: from action to contemplation, from calculation to understanding, and from output to experience.

Conclusions: Lessons from the Quest for Artificial Consciousness

The efforts to understand consciousness and create conscious machines have brought us interesting insights that have important implications for our understanding of human nature and its relationship with God, with others and with technology. The successes of AI have deeply impacted how we understand our own human nature and introduced some emerging ethical issues related to our responsibility in a world where an increasing number of decisions are being made by machines and where the limits between machines and humans are becoming increasingly difficult to define.

The core concept that can guide the evaluation of these successes, as well as their ethical implications, is human uniqueness and Imago Dei. In particular, we can consider the two main opposing views: the Augustinean view, which emphasizes intellectuality, and the relational view, which focuses on the subjective experience of relating to God and to others.

A first lesson is that the successes of AI have reinforced a dualist vision of humanness, where the two parts, which are now not body and soul, but body and information processing capabilities. Every time AI succeeds at an objective task that was previously only attainable to humans, according to this view, the machine and the human get closer. Moreover, given that the machine was able to achieve the target by using an electronic support and the human did it with a biological support, this reinforces the view that the support (i.e., the body) is irrelevant and replaceable.

However, the focus on objective outcome means that the most important aspects of consciousness—subjective experience—are largely ignored. This is complicated by the fact that the mechanisms that underlie RL and that guarantee its success in amenable problems, ones that can be objectively defined and where sufficient data and computation power are

available. This means that under favorable conditions, the machine will learn to pass any objective consciousness test, so it is unreasonable to believe that because the machine is able to pass a given test with the same metrics that a human would, then the machine is conscious.

Functionalist approaches to consciousness do not account for the key importance of subjective experience, which is by definition not objectively testable. In addition, as presented in this article, AI can potentially be trained to pass any test, which means that passing the test should not be equated with experiencing phenomenal consciousness. The emergence criterion has been presented as a useful basis to identify instances of AI where it is unreasonable to equate passing an objective test with having subjective experiences.

Our understanding of the complexity in algorithms is still developing, but we know already that there are some algorithms that are capable of developing long-range integrating patterns. This level of complexity would be necessary to sustain the information-integrative qualities that some definitions of consciousness focus on. However, the integration of information is not enough to generate artificial consciousness. Qualia could be completely independent from calculative power (in fact, it is very reasonable to assume they are), and embodiment is key to subjective experience, but they remain largely a mystery. We have no basis to assume they exist in current implementations of AI.

Even our more complex algorithms are not capable of generating insight, of integrating insight with common knowledge or judging whether there are potentially any biases that could lead any decisions to be faulty or unfair. It is necessary to take this into account when pondering the ethics of decision delegation, something that is becoming increasingly prevalent in our society. Only human beings have capacities for understanding and responsibility. Given this, it is necessary to support the development and application of interpretable ML to ML problems where decisions have relevant consequences for individuals: ML can be a decision support tool, but decisions should not be delegated to machines. In addition, it is necessary to realize that not focusing on subjectivity leads to making mistakes when judging consciousness in algorithms, as demonstrated in the examples of Sophia and LaMDA.

Not only is insight out of reach for our current ML, but so is true embodiment, which we can understand as the root of qualia, of phenomenal consciousness and of any subjective experience. The focus on subjective experience shifts what is relevant in our understanding of ourselves as human beings and as an image of God. This should lead us to de-emphasize intellectuality in favor of subjective experience, to value contemplation over action, to value the experience of love and affection, the enjoyment of the arts or the experience of emotions. Thus, the quest for artificial consciousness and AGI might not have yet brought us a key discovery, but it

has indeed opened our eyes to key issues regarding our own nature and our relationship to God, to others and to technology.

It is possible that, in the future, we find that the future does bring us artificial consciousness and AGI. What would the implications be? In that case, humans would have been achieved the most radical act of creativity there could be, becoming created co-creators and being joined by the machines, which would be a second generation of created co-creators. Only future will tell.

References

Agnati, Luigi Francesco, Diego Guidolin, Pietro Cortelli, Susanna Genedani, Camilo Cela-Conde, and Kjell Fuxe. 2012. "Neuronal Correlates to Consciousness. The 'Hall of Mirrors' Metaphor Describing Consciousness as an Epiphenomenon of Multiple Dynamic Mosaics of Cortical Functional Modules." *Brain Research* 1476:3–21.

Aleksander, Igor. 2008. "Machine Consciousness." *Scholarpedia* 3 (2): 4162.

Arulkumaran, Kai, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. "Deep Reinforcement Learning: A Brief Survey." *IEEE Signal Processing Magazine* 34 (6): 26–38.

Augustine, Saint. 1951. *De utilitate ieiunii: A Text with a Translation, Introduction and Commentary*, vol. 85. Washington, DC: Catholic University of America Press.

Baudot, Pierre. 2019. "Elements of qualitative cognition: An information topology perspective." *Physics of Life Reviews* 31:263–75.

Baum, Seth. 2017. "A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy." Global Catastrophic Risk Institute Working Paper: 17-1.

Bostrom, Nick. 2017. *Superintelligence*. Malakoff, France: Dunod.

Brunner, Emil. 2014. *The Christian Doctrine of Creation and Redemption: Dogmatics*, vol. II. Eugene, OR: Wipf and Stock Publishers.

Campolo, Alexander, and Kate Crawford. 2020. "Enchanted Determinism: Power Without Responsibility in Artificial Intelligence." *Engaging Science, Technology, and Society* 6:1–19.

Chalmers, David J. 2002. "Philosophy of Mind: Classical and Contemporary Readings."

———. 2009. "The Two-Dimensional Argument Against Materialism." In *The Character of Consciousness*, 141–91.

Chandrasekaran, Balakrishnan, John R. Josephson, and V. Richard Benjamins. 1999. "What are Ontologies, and Why Do We Need Them?" *IEEE Intelligent Systems and their Applications* 14 (1): 20–26.

De Aquino, Tomás, and Pietro Caramello. 1962. *Summa Theologiae*. Madrid: Editorial Católica.

Dehaene, Stanislas. 2014. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. New York: Penguin.

Deland, Jane S. 1999. "Images of God Through the Lens of Disability." *Journal of Religion, Disability & Health* 3 (2): 47–81.

Della Mirandola, Pico. 2012. *Oration on the Dignity of Man: A New Translation and Commentary*. Cambridge: Cambridge University Press.

Eiesland, Nancy L. 1994. *The Disabled God: Toward a Liberatory Theology of Disability*. Nashville, TN: Abingdon Press.

Froehlich, Thomas J. 2017. "A Not-So-Brief Account of Current Information Ethics: The Ethics of Ignorance, Missing Information, Misinformation, Disinformation and Other Forms of Deception or Incompetence." *BiD*, no. 39.

Gamez, David. 2018. *Human and Machine Consciousness*. Cambridge: Open Book Publishers.

Hajian, Sara, Francesco Bonchi, and Carlos Castillo. 2016. "Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining." Paper presented at Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Harnad, Stevan. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1–3): 335–46.

Hefner, Philip. 2019. "Biocultural Evolution and the Created Co-Creator." In *Science and Theology: The New Consonance*, edited by Ted Peters, 174–88. London: Routledge.

Hong, Joo-Wha, and Nathaniel Ming Curran. 2019. "Artificial Intelligence, Artists, and Art: Attitudes Toward Artwork Produced by Humans vs. Artificial Intelligence." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15 (2s): 1–16.

Kant, Immanuel. 2021. *Fundamentación de la metafísica de las costumbres*. Rome: Greenbooks editore.

Khayut, Ben, Lina Fabri, and Maya Avikhana. 2020. "Toward General AI: Consciousness Computational Modeling Under Uncertainty." Paper presented at 2020 International Conference on Mathematics and Computers in Science and Engineering (MACISE).

Koch, Christof, and Giulio Tononi. 2011. "Testing for Consciousness in Machines." *Scientific American Mind* 22 (4): 16–17.

Kurzweil, Ray. 2012. *How to Create a Mind: The Secret of Human Thought Revealed*. New York: Penguin.

Landgrebe, Jobst, and Barry Smith. 2019. "There is No Artificial General Intelligence." ArXiv Preprint arXiv:1906.05833.

Lumbreras, Sara 2017. "Strong Artificial Intelligence and Imago Hominis: The Risks of a Reductionist Definition of Human Nature." In *Issues in Science and Theology: Are We Special? Human Uniqueness in Science and Theology*, edited by Michael Fuller, Dirk Evers, Anne Runehov, and Knut-Willy Sæther, 157–68. Cham, Switzerland: Springer.

———. 2018. "Getting Ready for the Next Step: Merging Information Ethics and Roboethics—A Project in the Context of Marketing Ethics." *Information* 9 (8): 195–206.

———. 2020. "The Transcendent Within: How Our Own Biology Leads to Spirituality." In *Issues in Science and Theology: Nature—and Beyond. Transcendence and Immanence in Science and Theology*, edited by Michael Fuller, Dirk Evers, Anne Runehov, Kunt-Willy Sæther, and Bernard Michollet,187–97. Springer.

Luscombe, Richard. 2022. "Google Engineer Put on Leave After Saying AI Chatbot Has Become Sentient." *The Guardian*, June 12, 2022.

Manzotti, Riccardo. 2008. "Physical Foundations of Phenomenal Content in a Conscious Machine, Nokia Workshop on Machine Consciousness."

McGrath, Alister E. 2012. *Historical Theology: An Introduction to the History of Christian Thought*. Malden, MA: John Wiley & Sons.

Middleton, J. Richard. 1994. "The Liberating Image? Interpreting the Imago Dei in Context." *Christian Scholars Review* 24 (1): 8–25.

Molnar, Christoph. 2020. *Interpretable Machine Learning*. Lulu.com.

Newen, Albert, Leon De Bruin, and Shaun Gallagher. 2018. *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Pinar Saygin, Ayse, Ilyas Cicekli, and Varol Akman. 2000. "Turing Test: 50 Years Later." *Minds and Machines* 10 (4): 463–518.

Potember, Richard. 2017. *Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD*. Report. McLean, VA: Mitre Corp.

Raoult, A., and R. Yampolskiy. 2015. "Reviewing Tests for Machine Consciousness." *ResearchGate*.

Ravenscroft, Ian. 2005. *Philosophy of Mind: A Beginner's Guide*. Oxford: Oxford University Press.

Retto, Jesús. 2017. "Sophia, First Citizen Robot of the World." *ResearchGate*.

Ricoeur, Paul, and George Gingras. 1961. "'The Image of God' and the Epic of Man." *Cross-Currents* 11 (1): 37–50.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–15.

Russell, Stuart J. 2010. *Artificial Intelligence: A Modern Approach*. London: Pearson Education, Inc.

Searle, John R. 1982. "The Chinese Room Revisited." *Behavioral and Brain Sciences* 5 (2): 345–48.

Sharp, Lesley A. 2000. "The Commodification of the Body and Its Parts." *Annual Review of Anthropology* 29 (1): 287–328.

Skjuve, Marita, Asbjørn Følstad, Knut Inge Fostervold, et al. 2021. "My Chatbot Companion—A Study of Human-Chatbot Relationships." *International Journal of Human-Computer Studies* 149:102601.

Smith, Brian Cantwell. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: MIT Press.

Sutton, Richard S., and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Ta, Vivian, Caroline Griffith, Carolynn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero, and Alexia Loggarakis. 2020. "User Experiences of Social Support from Companion Chatbots in Everyday Contexts: Thematic Analysis." *Journal of Medical Internet Research* 22 (3): e16235.

Timmermans, Bert, and Axel Cleeremans. 2015. "How Can We Measure Awareness? An Overview of Current Methods." *Behavioural Methods in Consciousness Research*, edited by Morten Overgaard, 21–46. Oxford: Oxford University Press.

Tononi, Giulio, Melanie Boly, Marcello Massimini, and Christof Koch. 2016. "Integrated Information Theory: From Consciousness to its Physical Substrate." *Nature Reviews Neuroscience* 17 (7): 450–61.

Turing, Alan M. 2009. "Computing Machinery and Intelligence." In *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, edited by Robert Epstein, Gary Roberts, and Grace Beber, 23–65. New York: Springer.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin 2017. "Attention is All You Need." *Advances in Neural Information Processing Systems* 30:1–11.

Waltz, M., and K.S. Fu. 1965. "A Heuristic Approach to Reinforcement Learning Control Systems." *IEEE Transactions on Automatic Control* 10 (4): 390–98.

Washington, Anne L. 2018. "How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate." *The Colorado Technology Law Journal* 17:131–41.

Wheale, Nigel. 1991. "Recognising a 'Human-Thing': Cyborgs, Robots and Replicants in Philip K. Dick's 'Do Androids Dream Of Electric Sheep?' And Ridley Scott's 'Blade Runner.'" *Critical Survey* 1(3): 297–304.

Wu, Lijun, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu 2018. "A Study of Reinforcement Learning for Neural Machine Translation." ArXiv Preprint arXiv:1808.08866.

Zhou, Li, Jianfeng Gao, Di Li, and Heung-Yeung Shum 2020. "The Design and Implementation of Xiaoice, An Empathetic Social Chatbot." *Computational Linguistics* 46 (1): 53–93.