Aufklärung Revista de Filosofia

AUFKLÄRUNG, João Pessoa, v.11, n. Especial, Nov., 2024, p.11-24
DOI: https://doi.org/10.18012/arf.v11iEspecial.69910
Recebido: 30/05/2024 | Aceito: 30/10/2024

Can AI be trustworthy?[1]

[¿Puede la IA ser confiable?]

Rodrigo González *
Rodrigo GonzálezUniversidad de Chile, Chile

Abstract: In this paper, I address why some people distrust of Artificial Intelligence, and how this discipline can be trustworthy. Specifically, I describe the origins of such distrust, make a prognosis of the current state of uncertainty it provokes, and offer a key for a trustworthy AI. In the first section, I deal with why machines have not been considered trustworthy, and, more importantly, with the core of distrust of AI: as I contend here, there has been a shift from an academic debate about AI minds, due to skepticism about other minds, to fears on day-to-day life, mainly due to concerns of human replacement. In the second section, I make a prognosis about the current state of distrust AI provokes. Finally, I offer a key for a trustworthy AI based on explicability. As I argue, explicability via experts who have been certified by institutions are facilitators of a trustworthy AI.
Keywords: Artificial Intelligence; Distrust; Human replacement; Ethics of AI; Explicability.

Resumen: En este trabajo analizo por qué hay gente que desconfía en la IA, y cómo esta puede ser confiable. En específico, describo los orígenes de la desconfianza, hago un diagnóstico de la incertidumbre que pArovoca, y ofrezco una clave para una IA confiable. En la primera sección, trato con cómo las máquinas no han sido consideradas confiables, lo cual es la esencia de la desconfianza en la IA: tal como muestro, ha habido un cambio desde un debate académico acerca de las mentes-IA, a raíz del escepticismo de las otras mentes, a miedos cotidianos, principalmente basados en la preocupación por el reemplazo humano. En la segunda, hago un diagnóstico acerca de la incertidumbre que la IA provoca. Finalmente, ofrezco una clave para una IA confiable basada en la explicabilidad. Argumento que esta vía expertos certificados institucionalmente son factores facilitadores de una IA confiable.
Palabras clave: Inteligencia Artificial; Desconfianza; Reemplazo humano; Ética de la IA; Explicabilidad.

1 Introduction

We trust science (Hardwig, 1991; Hendriks et al., 2016). Trust is essential to science because we have a natural tendency to gain knowledge; for this reason, lay people trust scientists about things that they do not know. That is how lay

* *Rodrigo González Fernandez is an Associate Professor, Faculty of Philosophy and Humanities, Universidad de Chile. His research interests are Philosophy of Mind and Artificial Intelligence, Social Ontology and Epistemology. He has published a book entitled "Experimentos mentales y Filosofías de Sillón: Desafíos, Límites, Críticas", and articles in journals such as AI and Society, Anales del Seminario de Historia de la Filosofía, Isegoría, Revista de Filosofía, Aurora, Unisinos Journal of Philosophy, among others. E-mail: rodgonfer@gmail.com*

people gain understanding of the world on non-ordinary matters. On the other hand, scientists trust other scientists because the latter may be more knowledgeable than them. The whole process is part of the division of labor: those who do not know tend to trust those who do (Hendriks et al., 2016, p. 145-146), and this process is part of the epistemic inequality between the expert and the lay, which implies that the whole process has a social dimension (Barotta and Gronda, 2020, p. 81-82). How would humanity have made progress, otherwise? It seems that the core of human progress is related to how we have trusted science and, moreover, to how scientists who have become experts on some matters; usually, scientists as experts do not revise everything again and start from scratch, like in a Cartesian evil demon's dream. In fact, Tennie, Call and Tomasello (2009) hold that culture is cumulative, because it renders a ratchet effect. This process of accumulation allows the division of the epistemic labor, which in turn explains why we defer to experts. As part of this ratchet effect, scientists simply assume that certain bodies of knowledge are true, because others have been in a better position to gain knowledge and they have given the proper testimony; only then, scientists proceed to gain more knowledge about things that they ignore.

AI can be regarded as a technology, that is, as a branch of science that systematically applies knowledge for achieving practical goals. For this reason, AI is part of the contemporary sciences, despite being very novel. As a scientific discipline, the tradition considers that AI's birth was in 1956, when John McCarthy organized the Dartmouth Summer Research Project on Artificial Intelligence. It is usually thought that this year marks the consolidation of AI as a scientific discipline, because the meetings gathered all those who were interested in computer intelligence. Although such meetings were rather unsuccessful then, they gave rise to a new scientific community, one with a new sense of identity (Copeland, 2001, p. 8-9).

Here, I argue that part of the distrust of AI finds its origin in Descartes's criterion for mind and intelligence. I take this criterion as the origin of *the academic distrust of AI as a science*, which is a form of skepticism about other minds. However, there is a second more important sense on which I focus in this paper: distrust of AI as a technology, which will dramatically change our lives by replacing humans with machines. This is what I dub *the practical distrust of AI as a technology.* While the academic distrust has led to a debate on whether AI machines are minds, the practical distrust of AI is currently causing another heated debate, causing a state of uncertainty in the public about the future of humanity. This possibility leads to the research question of this paper: Can AI be trustworthy from a practical point of view? That is, can AI be trustworthy in the sense of being able to do what humans are interested to do?[2] This paper is divided into three sections and a conclusion.

In the first section, I identify Descartes' metaphysics as the source of the academic skeptical doubts about the possible existence of mind-machines. As the main focus of the paper is the practical distrust of AI, I briefly mention how Charles Babbage in the 19th century bet on the contrary view, i.e., that machines are trustworthy unlike humans, who are always prone to error. Interestingly, Descartes and Babbage's assumptions gave rise to an academic debate on whether AI can have mental states or, more bluntly, on whether AI machines are minds.

In the second section, I briefly concentrate on the prognosis of current AI. In particular, I identify the main factor that currently provokes uncertainty and anxiety

AUFKLÄRUNG, João Pessoa, v.11, n. Especial, Nov., 2024, p.11-24

about AI: the fear that humans will be dramatically replaced by AI machines. This factor is originated by *Babbage's replacement thesis*, because the British mathematician envisioned machines doing tasks that could surpass human intelligence in many domains. Turing, in turn, is the epitome of the replacement thesis, with his famous and controversial Imitation Game.

In the third section, I resort to Floridi's ethical approach to AI (Floridi, 2021b). In particular, I summarize the five principles that should rule AI, which have been inspired by the four principles of bioethics: beneficence, non-maleficience, autonomy, and justice (Beauchamp and Childress, 2012). Upon this basis, I concentrate on the fifth principle that Floridi lays down for an ethics of AI: explicability. Can AI be trustworthy if it is explicable? The answer is positive: by paying attention to how experts, who have been certified by institutions, restrict the aims of AI. As I conclude, then, Floridi's fifth principle is a key for a trustworthy AI.

2 The origins of distrust of AI

2.1 Descartes' distrust of machines: the indispensability of mind for intelligence

Famously, Descartes refers to the impossibility *in principle* that machines think, because, as corporeal things, they are limited due to physical mechanisms, all of which have limited outputs. These Cartesian metaphysical assumptions have an unseen consequence, one that anticipated a heated contemporary debate in the philosophy of mind. In one passage of the *Discourse on Method*, Descartes' emphasizes a criterion to distinguish between things that think and animals/machines: the use of language and intelligent action (AT 6, p. 56-57). I consider that Descartes' criterion for mind lays the foundation of *the academic distrust of AI:* some people do believe that there are good reasons to doubt that machines are minds. In fact, some philosophers believe that it is dubious that programmed machines are minds, for two reasons: i) such machines are unable to give unlimited and flexible linguistic outputs to problems that arise in different contexts; ii) such machines are incapable of intelligent actions, that is, their behavior is constrained to the working of programs, which cannot anticipate the right outputs for certain hard to anticipate contexts.[3]

Descartes is, then, the first philosopher who explicitly expresses distrust of the possible existence of machine-minds. The reaction is sharp amongst AI researchers, especially those who follow Turing's overturn of the question "Can a machine think?" via his imitation game in the 20th century (Turing, 1950; Turing, 1951; Turing et al., 1952). Take, for example, Weizenmbaum's Eliza (1976), and the wide range of chatbots that were created since 1950.

However, the other side of the coin also needs to be examined, that is, whether human minds are trustworthy in the sense of being able to fulfill the interests they were commended to. In fact, the very concept of intelligence has been under examination since the 19th century, and even before, which encouraged Turing to coin the term "machine intelligence." The next section precisely deals with how Babbage justifies the

possible existence of trustworthy machines in view of the problem of the calculation tables, and the need for accuracy and machine efficiency.

AUFKLÄRUNG, João Pessoa, v.11, n. Especial, Nov., 2024, p.11-24

2.2 Babbage: trustworthy machines and the replacement thesis

In the 19th century certain intellectuals considered machines more trustworthy than humans. That value is central to the industrial revolution, because at that time, it was usually thought that machines could be more efficient, accurate and tireless than humans. The story about the calculation tables depicts well in what sense humans were not considered sufficiently trustworthy according to the advocates of the industrial revolution. Before the construction of machines that could be accurate and replace human-made calculation tables, these were used for astronomy, construction, financial calculations, among other human activities.

It is worth noting that the calculation tables were hand made in four stages. First, their formulas were written by expert mathematicians. Second, a group of people, the so-called "computers," applied fixed formulas to calculate the figures by rote procedures. Third, these figures were written by hand in a manuscript. Fourth, there was a proofreading stage at which all the written figures were compared. Importantly, in all these stages human errors could be made, which meant that such tables, and even their errata, could be inaccurate. Worse yet, all these human errors caused deaths because of accidents, or there were important economic losses, all of which caused *uncertainty* (I will deal with this issue again below).

But there was an important effort to attack the negative effects of human error. Babbage, who attempted to construct a machine that could be sufficiently accurate, aimed to avoid the four stages at once. As a typical materialist, Babbage was the total opposite of Descartes. He thought that he could construct a machine to avoid human error by mechanizing calculation and mathematical reasoning. In other words, he thought that he could construct a machine that thinks like a tireless, free of errors, mathematician. This is the very genesis of a trustworthy machine in Babbage's terms: the Difference Engine.

After constructing the essential part of such an engine, many of its improvements encouraged Babbage to describe the machine with *psychological vocabulary*. In particular, certain terms alluded to the presence of a mind, for example, Babbage used terms such as "remember," "think," and "learn". However, he justified the use of psychological vocabulary in pure instrumental terms. Babbage's rationale was as follows:

> [with] the principle of successive carriages, it occurred to me that it might be possible to teach a mechanism to accomplish another *mental process*, namely—to *foresee*. This idea occurred to me in October 1834. It cost me much thought, but the principle was arrived at in a short time. As soon as that was attained, the next step was to teach the mechanism which could foresee to act upon that foresight (Babbage, 2010, p. 104-105, my emphasis).

Note that Babbage did not intend to create minds, which some AI researchers attempt to do nowadays. Instead, Babbage aimed to *replace* human labor with his machines. This emphasis on replacement gave birth to what I dub here as *Babbage's*

replacement thesis, i.e., the long-term project of replacing human labor with the aid of trustworthy AI machines and programs.

Babbage's replacement thesis underscores an economic factor, which is indeed related to a human fear: unemployment. When humans analyze whether AI machines and programs can replace them, most feel anxiety and fear of such possible replacements. Consequently, whether programmed machines are minds has turned into an academic debate; on the practical side, and as the replacement thesis allows to anticipate, the fear of unemployment was far more important when explaining daily life distrust of AI. A sign that the fear of human replacement is the correct diagnosis is the current wave of worry due to possible replacement of humans with machines. The dramatic news about people losing their jobs have become more and more prevalent these days.[4]

In the 20th century, Turing's philosophical project is the epitome of the replacement thesis. In fact, the core of Turing's project is the *replacement* of humans with machines capable of successfully performing well in the so-called imitation game. The next section elucidates how this game resorts to deception understood as X passes for Y so that the replacement of humans with programmed machines can take place.

2.3 Turing's philosophical project: the epitome of human replacement

Like Babbage, Turing does not refute explicitly Descartes' criterion of intelligence. Still, Turing goes a step forward, and takes the cue from the British mathematician. Put briefly, Turing's aim is to replace the knotty question "Can a machine think?". In fact, this question leads to the use of concepts such as "machine" and "intelligence", with all the negative consequences that ensue. The worst one is that analyzing such concepts may ultimately lead to a sort of Gallup poll on their common use. To avoid this problem, Turing presents a game, i.e., the imitation game, which helps replace the knotty question.

There are two stages in the game. The first stage describes a man in room A, a woman in room B, and interrogators of any sex. When the man replies to short questions of the interrogators, he attempts to *pass for the woman*. By contrast, she answers the questions sincerely; as a result, the interrogators attempt to guess the sex of those inside room A and B. The ideal outcome is that the man passes for a woman so that the interrogators get deceived and make the wrong identification.

Turing notes that, to avoid an easy identification due to the voices, all the questions and answers must be typed. The rounds of questions last 5 minutes or so; after that, the interrogators need to determine the sex of the participants. But, why does the sex of the participants is an important issue in the game? The tradition has not given importance to this detail, although it does have in virtue of Turing's functionalist view about machine intelligence, which is indeed related to Babbage's replacement thesis.

Turing's concept of "machine intelligence" needs to be grasped in terms of how a digital computer can *replace* humans, or how computers may pass for humans. On Turing's view, the intellectual capabilities of the man have to be distinguished from his sex, or the man's physical properties of his body and brain (Turing, 1950, p. 41). This means that, to deceive the interrogators, the man intends to *replace* the woman, in terms

of how she would have answered to all the questions. In fact, that is the core of the second stage of the imitation game.

This stage of the imitation game is described as two open questions by Turing: What would happen if a digital computer *played the role of man* in room A? In view of how the digital computer may *pass for a* woman, would the interrogators decide incorrectly as often as they would do when the man took part in the game? (Turing, 1950, p. 41). These two questions, Turing contends, replace the original knotty question "Can a machine think?" Moreover, in view of the game, it turns out to be absurd to examine whether a machine can think, which means Descartes' criterion for intelligence is countered. But, is that so?

AUFKLÄRUNG, João Pessoa, v.11, n. Especial, Nov., 2024, p.11-24

As a functionalist, Turing himself accepts a general possible objection that precedes the nine more specific objections to his imitation game. The general objection is as follows: "May not machines carry out something which ought to be described as thinking but *which is very different from what a man does*?" (Turing, 1950, p. 42, my emphasis).

Deception via imitation is the core of the imitation game (Saygin et al., 2003, p. 26). However, another component has not been acknowledged by the tradition either, that is, how the imitation of humans aims to make *replacement* possible. In the Imitation Game, the man and the computer are supposed to *replace* the woman, by passing like her. As I argue here, this emphasis on replacement has provoked a wave of distrust of AI, which in turn has caused a state of uncertainty. The next section elucidates the biological basis of such feelings, with emphasis on the current hype about AI.

3 Prognosis: the acute uncertainty about AI

Distrust of AI cannot be completely neutralized, on the practical side. This is explained by the fact that sometimes AI is designed by humans who can attempt to gain more power on other humans. These days, some algorithms can be created to gain control over other humans, which elucidates in what sense the human replacement affects people. Remember that some people are worried that they could lose their jobs by being replaced by AI machines. Take, for example, what happens with ChatGPT, which will similarly perform than humans. Such fears can lead to more and more uncertainty about AI, although the program themselves are not to blame. On the contrary, those humans who design the algorithms behind the programs are. Even so, the most pessimistic people believe that human autonomy, privacy and freedom are at stake because of the advent of AI machines. All this has caused a state of uncertainty in some humans.

Recent scientific studies have examined intolerance of uncertainty, an issue that can be associated with Babbage's intolerance of human error, and with Babbage's replacement thesis. Reviewing different studies, Tanovic et al. (2018) have focused on how high levels of uncertainty are not only related to different pathologies, such as depression, generalized anxiety disorder, social anxiety, panic disorder, and eating disorders (Brown et al., 2017), all of which cause anticipatory anxiety and maladaptive attempts to reduce it.

Interestingly, Tanovic at al., also show that such attempts are linked to symptoms such as worry, reassurance checking, and hypervigilance (Barlow, 2004; Krohne, 1993). Undeniably, intolerance of uncertainty is present in our daily life. In fact, some recent physiological measures show that there is a fundamental, evolutionary supported fear of the unknown which has biological bases (Carleton, 2016; Shihata et al., 2016).

Pathological worries about other cultures, which evince distrust, can be explained by such biological bases. Now, Artificial Intelligence, which may replace millions of jobs, seems to be causing uncertainty in the public, in a similar sense that Tanovic et al. describe, for example, by how worry is being caused. Indeed, the replacement of all human activities by AI machines, and especially of human jobs, can be related to the recent and future developments of AI. Then, what sort of uncertainty is producing distrust of AI nowadays?

Humanity is experiencing a feeling of uncertainty about the possible scenarios they may face due to the developments of AI, especially programs such as ChatGPT and other Large Language Models (LLM). Despite not being pathological, such feelings are causing worry and discomfort, which intensifies how threatening the whole situation feels and will feel like to the public. No certainty exists about the possible outcome and scenarios related to all the changes that future AI will bring about. In fact, while some AI researchers have anticipated the best possible scenarios in which AI will help humanity to achieve prosperity and peace, others have predicted scenarios in which AI causes recession and even the extinction of humanity. Can we anticipate and tackle the problems that AI may cause to humanity? In the next section, I offer a key for a trustworthy AI, especially in view of Babbage's replacement thesis, and whether AI can fulfill the human being's interests.

4 How AI can be trustworthy: explicability via experts

Given the origins and prognosis about the distrust of AI, I provide in this section a positive prospect for a trustworthy AI. This prospect is ruined by some people, who still believe unreasonable hypes about AI. Journalists keep on writing newspaper articles with all kind of stories about nightmarish risks due to the advent of AI. Moreover, and as I have shown in the previous sections, some people distrust of AI, even showing intolerance of an uncertain future. In fact, Babbage's replacement thesis seems to confirm all these irrational fears: some people believe that we will completely lose our autonomy, others fear that we are going to be enslaved by AI, and the most pessimistic ones even believe that humanity will be extinguished by the advent of ruthless machines. Some concerns are not only alarmist; they are also irresponsible, because many of them are based upon fears of the unknown, that is, of an uncertain future. Other concerns seem more justified, for example, those which are related to security and privacy issues.

Surely, AI will have a major impact on society. The important point to bear in mind is how humans take the impact, especially from the viewpoint of philosophy and ethics. An ethical framework for AI is necessary as a leading-edge technology. The reason is quite simple: by providing an ethical framework for AI, humans will have a set of principles with which they should behave, given the means-ends relationship of

AI technology. On the other hand, having such an ethical framework will encourage a better understanding of AI, especially in view of one of its essential principles: explicability. If humans sufficiently understand AI technology, the intolerance to an uncertain future can be tackled. Note that, although an ethical framework for AI does not guarantee a completely secure future, nor the complete disappearance of distrust of AI, it offers a tool to dispel most of the irrational fears based on uncertainty. However, the importance of the key principle of explicability needs some explanation.

According to Floridi (2021b, p. 8-15), there are five principles for an ethical AI, which have been posited given the four standard principles of bioethics. These principles are beneficence, non-maleficence, autonomy, and justice (Beauchamp and Childress, 2012). The application of the four principles results in these five ethical principles for AI: *beneficence,* as promoting well-being, preserving dignity and sustaining the planet; *non-maleficence*: privacy, security and "capability caution"; *autonomy:* the power to decide (to decide); *justice:* promoting prosperity, preserving solidarity, avoiding unfairness. Nevertheless, the fifth principle offered does not come from bioethics, and as I argue it turns out to be crucial for a trustworthy AI: *explicability.* This principle enables the other four principles through intelligibility and accountability (below I refer to their connection with intolerance of uncertainty).

In relation to *beneficence,* it is worth noting that AI technology must be beneficial to humanity. There is no consensus as to what "beneficence" means, though. While some people take it to mean the "well-being" of humanity and of all sentient beings, others characterize the principle as "common good" to benefit humanity or related to "human dignity" and "sustainability", meaning that AI should ensure the basic conditions of a good environment for future generations. How the principle of explicability has been posited by Floridi needs some explanation, indeed.

It is also important to emphasize that "do only good", or beneficence, does not entail "do no harm", which is *non-maleficence.* As Floridi remarks: "each one [the documents on the Ethics for AI] cautions against various negative consequences of overusing and misusing AI technologies (Cowls et al., 2018)" (Floridi, 2021b, p. 10). Prevention of infringements on personal privacy is better represented by the latter, for example. An arms race and the recursive self-improvement of AI have been of concern for the experts as well. Such concerns show that warning about the possible misuse of AI require avoiding harm, which is the very idea about an ethical framework for AI.

Autonomy is related to non-maleficence, because an ethical framework for AI must strike a balance between the decision-making power humanity must retain, and which power can be delegated to artificial agents. This is explained by the fact that humanity may delegate power to smart agencies. But, what is supposed to be the risk, then? An imbalance between artificial autonomy and human autonomy: what needs to be prevented is that the former affects the latter, in terms of how AI could eventually threaten human freedom. To avoid this risk, humans should choose how and why to delegate crucial decisions to AI systems. Consequently, it follows that human autonomy should be protected by adequately restricting the AI machines' decision-making power.

The fourth principle posited by this philosopher is *justice*, which is related to promoting prosperity, preserving solidarity, and avoiding unfairness. The core of this principle involves avoidance of unjust acts, such as discrimination. On the other hand, the principle of justice is associated with the need for shared benefit and shared

prosperity, which means that the development of AI should be in line with equal access to its benefits. For this reason, if humans are guided by the principle of justice, they should prevent threats to fair treatment and solidarity, for example, problems with social insurance and health care. However, an unseen problem remains, as Floridi points out: Are humans the patient receiving the treatment of AI, the doctor prescribing it, or both? This question leads to Floridi's emerging fifth principle: explicability.

Intelligibility and accountability are how *explicability* enables the other four principles to rule (Floridi, 2021b, p. 12). For him, humans can be either patients or doctors, because there should be no exclusive-or when AI technologies are implemented. A small fraction of humanity program AI doctors to do their work and, at the same time, a large fraction of humanity receive the AI treatment. It is worth noting that different terms have expressed this explicability principle: "transparency", "accountability", "understandable and interpretable AI". Importantly, such terms are to be applied by the experts. In fact, Floridi points out that,

> The addition of the principle of 'explicability,' incorporating both the epistemological sense of 'intelligibility' (as an answer to the question 'how does it work?') and in the ethical sense of 'accountability' (as an answer to the question 'who is responsible for the way it works?'), is the crucial missing piece of the AI ethics jigsaw. It complements the other four principles: for AI to be beneficent and non-maleficent, *we must be able to understand the good or harm it is actually doing to society, and in which ways* […] (Floridi, 2011b, p. 12, my emphasis).

"Transparency" is crucial for a trustworthy AI. In another context, there has been identified a direct relationship between trust and transparency, especially regarding the influence on the evaluation of political institutions (Hakhverdian and Mayne Source, 2012). Some limits to the positive effect derived from education have also been remarked, especially in low corruption countries (Frederiksen et al., 2016). Although the connection does not seem to be obvious for the issue of a trustworthy AI, it is: transparent and trustworthy AI institutions can guarantee that AI experts make the public gain trust in AI. Thus, explicability and transparency go hand in hand because AI experts can teach how to gain trust in AI. By doing so, AI experts can teach the lay how AI machines may act instead of humans, and who holds accountable in case of negative outcomes. That explains why transparency and accountability can offer the path to explicability, especially when AI trustworthy institutions are concerned.

It is worth noting that this process resembles a lesson learnt from philosophy of language and philosophy of science. In the discussion of meaning, language has been considered as a social tool whose adequate use is prescribed by the linguistic community, which must include experts (Putman, 1975). For example, the use of natural kind terms and the problem of their meaning. Note that the meaning of natural kind terms is essential to science, because as Bird and Tobin remark: "Scientific disciplines frequently divide the particulars they study into *kinds* and theorize about those kinds. To say that a kind is *natural* is to say that it corresponds to a grouping that reflects the structure of the natural world rather than the interests and actions of human beings" (Bird and Tobin, 2023, p. 1). This explains why experts need to fix the meaning of natural kind terms such as 'water,' 'tiger,' 'light,' and the like. In fact, experts are trustworthy when they are in the best possible position to gain knowledge, which is a

natural consequence of the division of labor. Only then experts are able to fix the meaning of natural kind terms. But who are the scientific experts and how they become what they are?

Putnam is not sufficiently clear about this point. A clarification is needed for the sake of a better understanding of the connection between trustworthiness and scientific knowledge and, especially, to prevent the typical skeptical argument against science, which may affect the trustworthiness of AI both as a science as well as a technology.

Scientific experts are *certified* by specific institutions. For example, universities[5] validate the expert's expertise within the community. In the context of natural kinds, the lay may learn from the expert the meaning of 'water', and the former starts applying the term to all kind of instances of water. By the same token, AI experts become experts when they are certified by certain specific institutions, only then experts can teach the lay when AI technology is trustworthy by means of revealing its explicability.

As the reader may have noted, I here endorse Putnam's essentialist theory to justify the relation between expertise and explicability. Likewise, I feel sympathy for how science avoid the typical maneuvers that scientific theories are never completely trustworthy. Are they in the end? Another element seems to be crucial for the explanation of how AI can be trustworthy: cooperation, which is also important in the process of acknowledging trustworthiness. In communities, members cooperate among them to gain knowledge (Origgi, 2004). In the case of AI, AI experts should cooperate with the lay so that the latter understand its explicability; on the other hand, the lay should cooperate with AI experts by trusting their expertise. For example, transparency, which is necessary for autonomy, can be guaranteed in relation to the power AI has over us *only* if the expert and the lay know who is accountable in case of undesired outcomes. If not, AI does not turn out to be trustworthy, as it should be.

In general, AI experts are certified by institutions, and with the aid of credentials (Smith at al., 2000).[6] Then, they can adequately analyze what desired and undesired outcomes may exist. If that is the case, experts can anticipate that some AI program can be a threat to autonomy. Only then the lay and other experts can be warned about the threat. As a result, there is a direct relationship between expertise and explicability, since the lay and other experts can be warned. In case non-experts had an opinion about the program being safe when it is not, such opinions would be totally discredited, and no one would rationally believe anything about the program being unsafe. On the contrary, if AI experts concluded that the program is safe, both the lay and other experts need to have the same knowledge. For all of them would be taught that the program is safe. Thus, the direct connection between expertise and trustworthiness is a key to guarantee that AI is trustworthy in case it is considered to be so by the proper experts.

In summary, to prevent the wave of distrust of AI that humanity is currently experiencing, with different degrees of intolerance of uncertainty, a key factor must be considered: how AI experts can teach the explicability of AI technology[7]. AI experts can promote transparency, accountability and intelligibility of AI systems, so that the irrational fears that fuel the hypes about AI as a grave and deadly threat can be neutralized. Only then the nightmarish scenarios will be treated as they deserve, namely, as part of the hypes about AI that must be subject to reasonable doubt.

5 Conclusion

AI is unfolding a drama for humanity nowadays. On the one hand, some people are excessively optimistic, and believe that AI will solve all the problems humanity faces such as the climate crisis, the political crisis of western democracy, and drudgery. On the other hand, there are those who distrust of AI, believing that this scientific discipline and technology has become a mere tool to control and oppress.

In this paper, I have adopted another possible stance. It is true that not trusting AI sometimes seems to be reasonable, especially if it is only used as a means for profit, or for human control by AI machines; however, I have also argued that a trustworthy AI, especially from the practical day-to-day life viewpoint, can blossom if an ethical framework for AI is correctly adopted.

In particular, I have shown the origins of distrust of AI, given the current prognosis, and give the key for a trustworthy AI. In function of Floridi's fifth principle of the framework for AI, that is, explicability, I have shown that transparency and accountability are two necessary elements for a trustworthy AI. Note that perfect transparency is not required; rather, experts need to grasp how AI systems work so that they can explain who is accountable when AI technology does not work as it should do.

To sum up: I have shown under what conditions a trustworthy AI may develop. I have shown that revealing the transparency and accountability of AI systems via experts can prevent irrational fears about undesired or fatal outcomes. Moreover, if AI experts educate laymen properly, the latter will not irrationally fear that AI puts humanity at risk. Or, as Turing would put it in relation to an objection to the imitation game, the educated opinion will finally be able to avoid having their "heads in the sand." Only then humanity will be set free from believing the hype, that is, from being filled with irrational fears about AI mostly based upon the unknown.

References

ANDERSON, C.; SINGER, M. The sensitive left and the impervious right: multilevel models. *Comparative Political Studies*, v. 41, p. 564-599, 2008.

BABBAGE, C. Excerpt from 'Passages from the Life of a Philosopher'. In: BABBAGE, C. *Babbage's calculating engines*: being a collection of papers relating to them; their history, and construction. Edited by Henry P. Babbage.Cambridge: Cambridge University Press, 2010. p. 83-288.

BAIER, A. *Trust*: the Tanner lectures of human values, 1991. <https://tannerlectures.utah.edu/_documents/a-to-z/b/baier92.pdf>.

BAROTTA, P.; GRONDA, R. Epistemic inequality and the grounds of trust in scientific experts. In: FABRIS, A. (Ed.) *Trust*: a philosophical approach. Cham: Springer, 2020. p. 81-94.

BEAUCHAMP, T.; CHILDRESS J. *Principles of biomedical ethics.* Oxford: Oxford University Press, 2021.

BIRD, A.; EMMA, T. Natural kinds. In: ZALTA, E.; NODELMAN, U. (Ed.) *The Stanford Encyclopedia of Philosophy*, 2023. <https://plato.stanford.edu/archives/spr2023/entries/natural-kinds>.

BROWN, M.; ROBINSON, L.; CAMPIONE, G. C.; WUENSCH, K.; HILDEBRANDT, T.; MICALI, N. Intolerance of uncertainty in eating disorders: a systematic review and

metaanalysis. *European Eating Disorders Review*, v. 25, p. 329-343, 2017.

CARLETON, R. N. Fear of the unknown: one fear to rule them all? *Journal of Anxiety Disorders*,v. 41, p. 5-21, 2016.

COPELAND, J. *Artificial intelligence*: a philosophical introduction. Malden, MA: Blackwell, 2001.

COWLS, J.; FLORIDI, L.: TADEO, M. The challenges and opportunities of ethical AI. *Artificially Intelligent*, 2018. <https://digitransglasgow.github.io/ArtificiallyIntelligent/contributions/04_Alan_Turing_Institute.html>

DESCARTES, R. *Oeuvres.* Edited by Charles Adam and Paul Tannery, new edn, edited by the CNRS, 11 vols. Paris: Vrin, 1974-1976. [all Descartes' works are cited as AT in this paper]

FLORIDI, L. Introduction: The importance of an ethics-first approach to the development of AI. In: FLORIDI, L. *Ethics, governance, and policies in artificial intelligence.* Cham: Springer, 2021a. p. 1-4.

FLORIDI, L. A unified framework of five principles for AI in society. In: FLORIDI, L. *Ethics, governance, and policies in Artificial Intelligence.* Cham: Springer, 2021b. p. 5-17.

FREDERIKSEN, M.; LARSEN, C; H. LOLLE. Education and trust: exploring the association across social relationships and nations. *Acta Sociologica* v. 59, p. 293-308, 2016.KLEIN, N. AI machines aren't 'hallucinating'. But their makers are. *The Guardian*, May 8, 2023. available at: <https://www.theguardian.com/commentisfree/2023/may/08/ai-machines-hallucinating-naomi-klein>.

HAKHVERDIAN, A.; MAYNE SOURCE, Q. Institutional trust, education, and corruption: a micro-macro interactive approach. *The Journal of Politics* v. 74, p. 739-750, 2012.

HARDIN, R. *Trust and trustworthiness*. New York: Russel Sage Foundation, 2002.

HARDWIG, R. The role of trust in knowledge. *The Journal of Philosophy,* v. 88, p. 693-708, 1991.

HENDRIKS, F., KIENHUES, D.: R. BROMME. Trust in science and the science of trust. In: BLÖBAUM, B. (Ed.) *Trust and communication in a digitalized world*. Cham: Springer, 2016. p. 143-160.

MCLEOD, C. Trust. In: ZALTA, E.; NODELMAN, U. (Ed.) *The Stanford Encyclopedia of Philosophy*, 2021. <https://plato.stanford.edu/archives/fall2021/entries/trust/>.

MURRAY, S. The frame problem. In: ZALTA, E. (Ed.) *The Stanford Encyclopedia of Philosophy*, 2016. <https://plato.stanford.edu/archives/spr2016/entries/frame-problem/>.

ORIGGI, G. Is trust an epistemological notion? *Episteme* v. 1, p. 61-72, 2004.

PUTNAM, H. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, v. 7, p. 215-271, 1975.

SAYGIN, A. P.; CICKELI, I.; AKMAN V. Turing Test: 50 years later. In: MOOR, J. (Ed.) *The Turing test*: the elusive standard of artificial intelligence. Dordrecht: Kluwer Academic Publishers, 2003. p. 23-78.

SEARLE, J. *Making the social world*: the structure of human civilization. Oxford: Oxford University Press, 2010.

SHANAHAN, M. The frame problem. In: ZALTA, E. (Ed.) *The Stanford Encyclopedia of Philosophy*, 2016. <https://plato.stanford.edu/archives/spr2016/entries/frame-problem>.

SHIHATA, S.; MCEVOY, P; MULLAN, B.; CARLETON, R. Intolerance of uncertainty in emotional disorders: What uncertainties remain? *Journal of Anxiety Disorders,* v. 41, p. 115-124, 2016.

SMITH, B. *The promise of Artificial Intelligence*: reckoning and judgement. Cambridge, MA: The MIT Press, 2019.

SMITH, B.; LODDO, O. G.; LORINI, G. On credentials. *Journal of Social Ontology*, v. 6, p. 47-67, 2020. SWADE, D. *The difference engine*: Charles Babbage and the quest to build the first computer. London: Penguin Books, 2002.

TANOVIC, E., GEE, D. G.; JOORMANN, J. Intolerance of uncertainty: neural and psychophysiological correlates of the perception of uncertainty as threatening. *Clinical Psychology Review*, n. 60, p. 87-99, 2018.

TENNIE, C.; CALL, J.; TOMASELLO, M. Ratchering up the ratchet: on the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, v.364, p.2405-2415, 2009. <http://doi.org/10.1098/rstb.2009.0052>.

TOMASELLO, M. *Why we cooperate*. Cambridge, MA: The MIT Press, 2009.

TOMASELLO, M. *Being human*: a theory of ontogeny. Cambridge, MA: Harvard University Press, 2019.

TURING, A. Computing intelligence and machinery. *Mind*,n. 59, p. 433-460, 1950.

TURING, A. Can digital computers think? In: SHIEBER, S. (Ed.) *The Turing test*: verbal behavior as the hallmark of intelligence. Cambridge, MA: The MIT Press, 1951. p. 111-116.

TURING, A.; BRAITHWAITE, R.; JEFFERSON, G.; M. NEWMAN. Can automatic calculating machines be said to think? In: COPELAND, J. (Ed.) *The essential Turing*: seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life plus the secrets of enigma. Oxford: Oxford University Press, 1952. p. 487-506.

WEIZENBAUM, J.. *Computer power and human reason*: from judgement to calculation. San Francisco: W. H. Freeman & Company, 1976.

WILLIAMS, D. Yoval Noah Harari argues that AI has hacked the operating system of human civilization, 2023. <https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation>.

Notas

1 This research is part of the 1230128 ANID FONDECYT project: *Desconfianza: un factor causal de las crisis institucionales serleanas*. I would like to express my deepest gratitude to Marco Ruffino, Felipe Morales, Nara Figueiredo, César Meurer, Ludovic Soutif and Felipe Álvarez. I had the chance to discuss several points of this paper with them at the 3d Meeting on Cognition and Language, Campinas, Brasil.

2 I adopt Hardin's view here, i.e., trustworthiness in terms of the encapsulated view of trust (Hardin, 2002).

3 It is worth mentioning that the inability of machines to fit into different contexts gave rise to the so-called frame problem in the 20th century, that is, the difficulties a machine has to deal with when different contexts require different parameters, all of which cannot been anticipated by the machine (See Murray, 2016).

4 Take for example this recent world economic forum report: https://open.substack.com/pub/aisciencenews/p/ai-will-replace-85-million-jobs-by?utm_campaign=post&utm_medium=web.

5 Surely, there are many institutions that certify experts. I here provide the example of universities because they are paradigmatic for the problem of expertise.

6 Someone may object that on occasions "AI experts" are like that because a given community has considered them to be so. However, what I defend here is the normative side of expertise. In particular, some people may believe that an AI expert is an expert, but she may not be for different reasons (for example, she may have impressed others about things that she "knows"). Obviously, in that case she would only count as an expert until it is proved that she is not. Consider this: a fake doctor is considered to be a doctor until it is proved that her diplomas are fake. That is the reason why fake experts use counterfeit certifications.

7 In another context, education has been identified as a key factor of institutional trust (Anderson and Singer, 2008, p. 22-23). Not only the use of natural language, with the development of essential cognitive tools, can boost forms of cooperation within the community (Tomasello, 2009; Tomasello, 2019).

AUFKLÄRUNG, João Pessoa, v.11, n. Especial, Nov., 2024, p.11-24