WHAT SHOULD WE REASONABLY EXPECT FROM ARTIFICIAL INTELLIGENCE?

LEONARDO PARENTONI
Tenured Law Professor
at the Federal University of Minas Gerais – UFMG

SUMMARY – 1. Introduction. – 2. The many ways to provide AI-based products and services. – 3. Concerns about AI accuracy. – 4. Different levels of AI-based systems interference in human decision-making. – 5. What should we reasonably expect from AI? – 5.1. Expected accuracy rate. – 5.2. Transparency/Explainability. – 5.3. Ban on AI or mandatory human intervention ("human in the loop"). – 6. Connecting this study with the international debate. – 7. Conclusion. – 8. References.

1. – Artificial Intelligence (or just AI) is one of the most pervasive and cutting-edge technologies of our time. It is already presented in a variety of sectors, such as agriculture, industry, commerce, education, professional services, smart cities, cyber defense, and so forth[1]. However, what is AI and what should we humans reasonably expect from it? "*That's an easy question to ask and a hard one to answer*", points out the literature[2].

The first step to properly answer that question is acknowledging that AI is *not* a single, monolithic concept. On the contrary, AI-based products and services embrace a wide variety of sector-specific applications with *different purposes, accuracy, and risks*. There is no one-size-fits-all definition of AI[3]. For instance, internationally the OECD considers that "*an AI system is a machine-based system that can, for a given set of human-defined objectives, make*

[1] OECD – Organization for Economic Co-Operation and Development. *OECD AI Principles overview*. Available at: <https://oecd.ai/en/ai-principles>. Access: 20 Jun. 2022, 3: «Artificial Intelligence (AI) is a general-purpose technology that has the potential to improve the welfare and well-being of people, to contribute to positive sustainable global economic activity, to increase innovation and productivity, and to help respond to key global challenges. It is deployed in many sectors ranging from production, finance and transport to healthcare and security».

[2] J. KAPLAN, *Artificial Intelligence: What everyone needs to know*. Oxford, 2016, 1.

[3] S. J. RUSSELL – P. NORVIG. *Artificial Intelligence*: *A Modern Approach*, 4. ed., London, 2022, 20. «The methods used [to define AI] are necessarily different: the pursuit of human-like intelligence must be in part an empirical science related to psychology, involving observations and hypotheses about actual human behavior and thought processes; a rationalist approach, on the other hand, involves a combination of mathematics and engineering, and connects to statistics, control theory, and economics».

predictions, recommendations, or decisions influencing real or virtual environments"[4]. In the legal field it is *"best understood as a set of techniques aimed at approximating some aspect of human or animal cognition using machines"*[5] or as *"machines that are capable of performing tasks that, if performed by a human, would be said to require intelligence."*[6] On the one hand, psychology authors refer to intelligence as *"a biopsychological potential to process information that can be activated in a cultural setting to solve problems or create products that are of value in a culture."*[7] Computer scientists, on the other hand, tend to focus their attention on each specific AI sub-field[8], such as expert systems, machine learning, neural networks, robotics, computer vision, and natural language processing. Finally, the June 2022 ISO/IEC 22989 proposed an international standard for "artificial intelligence concepts and terminology"[9].

No matter the field, the term "artificial intelligence" is misleading because it directly associates algorithmic processes with a simulation of human intelligence. Neuroscientists strongly refuse that kind of association[10]. More than just a terminological problem, the expression artificial intelligence deviates us from what really matters and brings

[4] OECD – Organization for Economic Co-Operation and Development. *OECD AI Principles overview*. Available at: <https://oecd.ai/en/ai-principles>. Access: 20 Jun. 2022, 7.

[5] R. CALO, *Artificial Intelligence Policy: A Primer and Roadmap*, in *University of Washington Research Paper*, 4 ss.

[6] M.U. SCHERER, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies*, in *Harvard Journal of Law & Technology*, Cambridge, v. 29, n. 02, 2016, 362.

[7] H. GARDNER, *Intelligence Reframed*: *Multiple Intelligences for the 21st Century*, New York, 1999, 33.

[8] For a detailed description of AI evolution, see: N.J. NILSSON, *The Quest for Artificial Intelligence*: *A History of Ideas and Achievements*, Cambridge, 2010.

For a list of some AI sub-fields, see: J. VIJIPRIYA, *et al*, *A Review on Significance of Sub Fields in Artificial Intelligence*, in *International Journal of latest trends in Engineering and Technology – IJLTET,* New Delhi, v. 06, n. 03, 2016.

[9] The transversal legal instrument to regulate the design, development and use of artificial intelligence systems – CAI, a Council of Europe work in progress, will also propose some standards, focused on human rights and environmental issues.

[10] For instance: H.L. DREYFUS, *Alchemy and Artificial Intelligence*, *Rand Corporation Report Papers*, 62. «Thus, the strong claim that every processable form of information can be processed by a digital computer is misleading».

Also: M.A.L. NICOLELIS – R. CICUREL, *The Relativistic Brain*: *How it Works and why it cannot be simulated by a Turing Machine*, Montreux, 2015.

A. DAMASIO, *Self Comes to Mind*: *Constructing the Conscious Brain*, New York, 2010, 43: «(…) the real problem of these metaphors [of human brains and machines] comes from their neglect of the fundamentally different statuses of the material components of living organisms and engineered machines».

nonsensical questions to the debate. Jerry Kaplan summarized the problem through a clever analogy:

> *«To better understand how the aspirational connection between machine and human intelligence clouds and colors our understanding of this important technology, imagine the confusion and controversy that powered flight might have suffered if airplanes were described from the start as "artificial birds". This nomenclature would invite distracting comparisons between aviation and avians, sparking philosophical debates as to whether airplanes can really be said to "fly" as birds do, or merely simulate flying. (…) If this misplaced framing had persisted, there might have been conferences of experts and pundits worrying about what will happen when planes learn to make nests, develop the ability to design and build their own progeny, forage for fuel to feed their young, and so on*»[11].

Scientific expressions better than AI would be "analytical computing"[12] or "machine behaviour"[13]. However, since "artificial intelligence (AI)" has become the leading term, worldwide, this study will adopt it. Another preliminary disclaimer is that this article is *not* focused on any specific AI sub-field in any given market sector. The purpose herein is to *take a step back* and dig deeper into a structural preliminary question of paramount importance to regulators, developers, and customers: *what should we reasonably expect from AI?*

In order to properly answer that question this article proceeds as follows: Section 2 describes the multiple ways of providing AI-based products and services, to contextualize how this technology is in the field, emphasizing the importance of a case-by-case analysis; Section 3 highlights that the scientific literature is deeply concerned about the accuracy rate in AI systems, sometimes implying that these systems should surpass human capabilities, no matter the context; Section 4 provides the author's own 3-level categorization of AI interference in the human decision-making process; Section 5 discusses the original misalignment between what some people expect from AI and what this technology can actually deliver, providing the author's criteria to set what we should reasonably expect from AI in each context, based on the purpose of using that technology, the level

[11] J. KAPLAN, *Artificial Intelligence*: *What everyone needs to know*, Oxford, 2016, 16 ss.

[12] ID., *op. cit.*, p. 17.

[13] I. RAHWAN, *et al*, *Machine behaviour*, *Nature*, v. 568, 481: «(…) we now catalogue examples of machine behaviour at the three scales of inquiry: individual machines, collectives of machines and groups of machines embedded in a social environment with groups of humans in hybrid or heterogeneous systems. Individual machine behaviour emphasizes the study of the algorithm itself, collective machine behaviour emphasizes the study of interactions between machines and hybrid human–machine behaviour emphasizes the study of interactions between machines and humans».

of AI interference in human decision-making, accuracy rates, risk analysis and transparency; finally, Section 6 contextualizes how this way of reasoning is already present in many national and international regulatory initiatives.

2. – *"Different types of AI systems raise different policy opportunities and challenges"*[14].

Not only does AI encompass a wide variety of market sectors but there is also a myriad of ways to develop and introduce the same application in each sector, depending on the developers' and retailers' strategy[15]. *Each one of them fulfills different purposes, accuracy, and risks*. These differences must be considered when answering the question: what should we reasonably expect from AI?

One of the fundamental differentiations concerns *embodied* versus *bodiless* AI. Embodied AI applications are those in which the artificial intelligence system is indissociably part of a corporeal product, such as industrial equipment, and autonomous cars. To be fully operational, this kind of application needs a predetermined physical structure[16]. Embodied AI is usually called a *robot*[17]. There are important subdivisions though. A robot that resembles a human being is called a *humanoid* or *android*, while

[14] OECD – Organization for Economic Co-Operation and Development. *OECD Framework for the Classification of AI Systems*. Available at: <https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en>. Access: 20 Jun. 2022. p. 16.

[15] H. NISSENBAUM, *How Computer Systems Embody Values*, In *Computer*, New York, v. 34, n. 03, 2001, 120: «Values affect the shape of technologies. Briefly, the values that systems and devices embody are not simply a function of their objective shapes. We must also study the complex interplay between the system or device, those who built it, what they had in mind, its conditions of use, and the natural, cultural, social, and political context in which it is embedded—for all these factors may feature in an account of the values embodied in it».

[16] As developed in classical writings, for instance: H. MORAVEC, *Robot: Mere Machine to Transcendent Mind*, Oxford, 1999. Preface. «This book has been brewing for nearly fifty years, since preschool adventures with a mechanical construction set implanted the consuming notion that inanimate parts could be assembled into animate beings. The brew bubbled over in an article in 1978, a book in 1988, and this work in 1998».

[17] S. J. RUSSELL – P. NORVIG, *Artificial Intelligence*: *A Modern Approach*, 3. ed., New Jersey, 2010, 971: «Robots are physical agents that perform tasks by manipulating the physical world».

S. NOLFI, *Behavioral and Cognitive Robotics*: *An adaptive perspective*, 2021, 8: «(…) we can define a robot as an artificial system that: (i) has a physical body that includes actuators, sensors, and a brain, (ii) is situated in a physical environment and eventually in a social environment including other robots and/or humans, and (iii) exhibits a behavior performing a function».

robots designed to have specific social interactions and provoke human emotions are called *social robots*[18] (most social robots mimic dolls and house animals). Authors are advocating that social robots should not be treated like ordinary property, but as "part of the family" instead, such as the pets, since they can build deep bonds to humans (especially children), profoundly affecting their emotions and social interactions. The same line of reasoning considers that social robots can be *victims* of abuse[19].

Conversely, bodiless AI is not bound to any specific physical structure at least on the customer's side. This kind of application can run simultaneously on many devices, with almost the same accuracy rate and reach a greater audience. A good example is cloud-based services.

The distinction matters because a bug in an embodied AI will probably cause only local damage, while a bug in a bodiless system can cause worldwide problems, depending on how the product or service was managed. On top of that, a bug in a social robot can cause long-lasting psychological problems and compromise social interaction, while a bug in industrial equipment inside an ordinary factory would hardly cause the same kind of damage. Therefore, the *purpose* of using each system, the *expected accuracy* and *transparency*, as well as the acceptable *risks* greatly differ based on how the product or service was provided.

Many other classifications exist but bringing a thorough description of them is not the purpose of this article. Suffice to say that how each system was designed and delivered is *one of the core factors* to be considered when properly assessing what we should reasonably expect from AI. Moreover, the analysis should be run *on a case-by-case basis*, considering the specifics of each situation.

[18] K. DARLING, *Extending Legal Protection to Social Robots*: *The effects of anthropomorphism, empathy, and violent behavior towards robotic objects*, in: *WeRobot*, 2012, 2: «A social robot is a physically embodied, autonomous agent that communicates and interacts with humans on a social level. (..) Social robots communicate through social cues, display adaptive learning behavior, and mimic various emotional states. Our interactions with them follow social behavior patterns and are designed to encourage emotional relationships. Examples of early social robots include interactive robotic toys like Sony's AIBO dog and Innovo Labs' robotic dinosaur Pleo (…)».

[19] *Op. cit.* p. 16: «This section proposes that abuse protection for social robots could follow the analogy of our animal abuse protection laws. Despite the fact that the exact underpinnings of animal abuse protection are contested and many do not match the reasons we might protect robots, there are both psychological and philosophical parallels».

See also the American Society for the Prevention of Cruelty to Robots, founded in 1999: http://www.aspcr.com/index.html

3. – *«Ask commentators why there is so much "hype" surrounding machine learning, and the response will often be a variant of one word – accuracy»*[20]

As the quotation above suggests, there is a widespread concern about the accuracy rate of AI, which roughly speaking means the level of precision an AI system can provide when compared to human standards[21]. The better the results provided by the system, the higher its accuracy rate. Some studies argue that the *observed* accuracy directly influences human trust in AI systems, even when the *actual* accuracy is lower[22]. Other studies point out that machine learning models may be *less reliable* than they appear[23].

Therefore, accuracy is a key factor in the scientific literature when assessing the efficiency of an AI system. Most papers on that subject, irrespective of the field of expertise or the purpose of the paper usually address accuracy, if not as the core argument, but at least as part of the text. This can be seen in fields such as health[24], botany[25], stock markets[26], military applications[27] and law[28], among many others.

[20] D. LEHR – P. OHM, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, in *University of California Davis Law Review*, Davis, v. 51, n. 02, 2017, 710.

[21] Explaining the concept of accuracy in classification software's: W. MEIRA JUNIOR – M.J. ZAKI, *Data Mining and Machine Learning*: *Fundamental Concepts and Algorithms*, 2. ed., Cambridge, 2020, 547: «The accuracy of a classifier is the fraction of correct predictions over the testing set. (…) Accuracy gives an estimate of the probability of a correct prediction; thus, the higher the accuracy, the better the classifier».

[22] M. YIN – J.W. VAUGHAN – H. WALLACH, *Understanding the Effect of Accuracy on Trust in Machine Learning Models*, in *Conference on Human Factors in Computing Systems*, Glasgow, 2019, 1: «We find that people's trust in a model is affected by both its stated accuracy and its observed accuracy, and that the effect of stated accuracy can change depending on the observed accuracy».

[23] S. LAPUSCHKIN, *et al*, *Unmasking Clever Hans predictors and assessing what machines really learn*, in *Nature Communications*, v. 10, 2019, 1: «Current learning machines have successfully solved hard application problems, reaching high accuracy and displaying seemingly intelligent behavior. (…) our work intends to add a voice of caution to the ongoing excitement about machine intelligence and pledges to evaluate and judge some of these recent successes in a more nuanced manner».

[24] M.K. SANTOS, *et al*, *Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine*, in *Radiologia Brasileira*, São Paulo, v. 52, n. 06, 2019, 387: «We have observed an exponential increase in the number of exams performed, subspecialization of medical fields, and increases in accuracy of the various imaging methods (…)».

B. NISTAL-NUÑO, *Artificial intelligence forecasting mortality at an intensive care unit and comparison to a logistic regression system*, in *Einstein*, São Paulo, v. 19, 2021.

[25] A.B. SCHIKOWSKI, *Modeling of stem form and volume through machine learning*, in *Agrarian Sciences – Anais da Academia Brasileira de Ciências*, Rio de Janeiro, v. 90, n. 04, 2018, 3389: «The objective was analyzing the accuracy of machine learning (ML) techniques in relation to a volumetric model and a taper function for acácia negra».

Accuracy is surely relevant and concerns about it are rooted in scientific literature. The problem is that an *excessive* focus on accuracy can lead to the false premise that AI should *always* surpass human capabilities no matter the context. And that is not true…

Indeed, there are many contexts in which AI should *not* necessarily beat human standards or even get close to them. In these cases, AI can play a significant role just by replacing human labor, even at the cost of a substantial decrease in accuracy. Gains in other factors, such as risk prevention or transparency can compensate for the loss of precision. Therefore, the *false* premise that AI must beat humans ends up undermining or even disregarding the importance of other relevant factors. Some studies[29] are already casting light on that point. Some accuracy limitations may be acceptable and eventually unavoidable. This does not necessarily undermine the suitability of AI.

Thus, each *purpose* to the use of an AI system (also considering how the system was provided) defines the *expected accuracy, transparency, and risk prevention*. It is a *mix* of these and other factors, on a *case-by-case basis*, that

[26] A. PATEL – D. PATEL – S. YADAV, *Prediction of Stock Market Using Artificial Intelligence*, Available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3871022>. Access: Feb. 16. 2022., 1: «In this model we will introduce and review more a possible way to predict stock movements with high accuracy».

[27] G.M. LIMA FILHO, *et al*, *Decision Support System for Unmanned Combat Air Vehicle in Beyond Visual Range Air Combat Based on Artificial Neural Networks*, in *Journal of Aerospace Technology and Management*, São José dos Campos, v. 13, 2021, 1: «In a beyond visual range (BVR) air combat, one of the challenges is identifying the best time to launch a missile, which is a decision that must be made quickly. (…) The ANN was trained with a data set with 1093 registered shoots in military exercises, and it shows 78.0% accuracy with the cross-validation procedure».

[28] D.L. BURK, *Algorithmic Legal Metrics*, Available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3537337>. Access: Feb. 16. 2022., 1: «Specifically, this paper shows how the problematic social effects of algorithmic legal metrics extend far beyond the concerns about accuracy that have thus far dominated critiques of such metrics».

[29] M.M. MALIK, *A Hierarchy of Limitations in Machine Learning*, in *Berkman Klein Center for Internet & Society Research Paper*, 2020. Available at: <https://arxiv.org/abs/2002.05193>. Access: Mar. 10. 2022, 1: «Machine learning has focused on the usefulness of probability models for prediction in social systems, but is only now coming to grips with the ways in which these models are wrong—and the consequences of those shortcomings. This paper attempts a comprehensive, structured overview of the specific conceptual, procedural, and statistical limitations of models in machine learning when applied to society. Machine learning modelers themselves can use the described hierarchy to identify possible failure points and think through how to address them, and consumers of machine learning models can know what to question when confronted with the decision about if, where, and how to apply machine learning».

should be considered when assessing the suitability of an AI system[30]. Accuracy alone is just one piece in the puzzle. Having that in mind, the following sections will describe other technical and legal factors that should also be weighed to assess what we should reasonably expect from AI.

4. – Many studies show that technological evolution can lead humankind to a new paradigm[31]. In this context, AI-based systems will have an increasing influence in the human decision-making process. This certainly brings new opportunities as well as risks[32]. This section will describe the *author's own 3-level categorization* of AI interference in the human decision-making process. This categorization is one of the core factors to be weighed when assessing what we should reasonably expect from AI.

Indeed, there are various categorizations concerning the level of AI-based systems interference in human decision-making. For instance, the National Highway Traffic Safety Administration – NHTSA, the agency for transportation safety in the US, has initially split autonomous vehicles into 5 levels, from 0 (no automation at all) to 4 (full self-driving)[33]. The NHTSA used the term *"automation"*, although some authors advocate that

[30] Who may (or should) do the assessment is a topic outside the scope of this study.

[31] For instance: J.M. BALKIN, *The Three Laws of Robotics in the Age of Big Data*, in *Yale Law School Research Paper n. 592*, 2017, 2 ss.: «Indeed, we are rapidly moving from the age of the Internet to the Algorithmic Society. We will soon look back on the digital age as the precursor to the Algorithmic Society. What do I mean by the Algorithmic Society? I mean a society organized around social and economic decision making by algorithms, robots, and AI agents; who not only make the decisions but also, in some cases, carry them out».

See also: A.C.M. CANSIAN, *Aspectos Jurídicos Relevantes da Internet das Coisas (IoT): Segurança e Proteção de Dados*, 2021, 136 ss., Tese (Doutorado em Direito Comercial) – Faculdade de Direito, São Paulo, 2021, 95: «If machines were up to now supporting actors for humans, they will not be anymore, they will become protagonists in human relationships and their unfolding in different scenarios, replacing human decision-making power and performing complex tasks, hitherto unthinkable for a computer».

[32] A. FÜGENER, *et al*, *Will Humans-in-The-Loop Become Borgs*? *Merits and Pitfalls of Working with AI*, available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3879937 >. Access: Feb. 16. 2022, 1: «(...) we claim that humans interacting with AI behave like 'Borgs', that is, cyborg creatures with strong individual performance but no human individuality».

[33] Detailed report Available at: <http://www.nhtsa.gov/nhtsa/av/pdf/Federal_Automated_Vehicles_Policy.pdf>. Access: Feb. 18. 2022.

For an analysis of theses 6 levels, see: C.R.P. LIMA, *Sistemas de Responsabilidade Civil para Carros Autônomos*, 2020, Tese (Professor Titular de Direito Comercial), Ribeirão Preto, 2020, 130 ss.

"autonomy"[34] would be more technical. In this article both expressions are used indistinctly.

The proposal herein is simpler and more intuitive. Besides, it is more useful as a reasoning tool than sector-based categorizations, since it applies to any player in any market sector, irrespective of the kind of system at stake, and the purpose of using the technology. It splits AI systems into *3 levels of interference* in the human decision-making process: 1) *task-automation auxiliary systems*; 2) *advisory systems*; and 3) *full decision-making*.

Since reality is much more complex than theory, of course, there will be hard cases in which the system at stake can be situated in a grey area between these levels. Even so, that categorization will have already served its purpose just by shedding light on most situations. Moreover, it is just one of the core factors to be considered. Bearing that in mind, let's briefly look at each level.

Task-automation auxiliary systems are the lowest level of automation. They work merely as a tool to provide information to a human user (serving as a digital catalog) or to execute tasks based on commands previously defined by that user. It is always the human being who will make all the decisions on their own. Moreover, these systems do not take the initiative of working out of the blue. They only work on the user's demand (passive functioning). These systems can eventually perform based on the user's decision, but they do *not* replace *any* stage of the human decision-making process. Even when this kind of system seems to be doing something on its own, such as booking an appointment or making a purchase, these actions have been previously set by the human user. Voice assistants[35] such as Siri

[34] J. CHERRY – D. JOHNSON, *Maintaining command and control (C2) of lethal autonomous weapon Systems: Legal and policy considerations*, in *Southwestern Journal of International Law*, Los Angeles, v. 27, n. 01, 2021, 4: «*Autonomy* is the ability of a machine to perform a task without human input. It is distinct from *automation*, which is simply using a machine to perform a particular process, while autonomy describes a system capable of operating independently for some period without direct human intervention (…) There are three basic dimensions of autonomy: the type of task the machine is performing; the relationship of the human to the machine while performing that task; and the sophistication of the machine's decision-making when performing the task. These dimensions are independent, and a machine can be 'more autonomous' by increasing the amount of autonomy along any of these spectrums. There are degrees of autonomy within these tasks, or dimensions, that dictate the human-machine relationship».

[35] OECD – Organization for Economic Co-Operation and Development. *OECD Framework for the Classification of AI Systems*. Available at: <https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en>. Access: 20 Jun. 2022, 55: «Automated voice assistant: Uses Natural Language Processing (NLP) to match user text or voice input to executable commands. Many continually learn using AI techniques including machine learning.

and Alexa usually fit inside this group. For instance, the human user can ask the assistant to "show pizza places close to my home". The AI will then show a list of results, probably with public rates/stars pointing to the best options. Then it's up to the human user to assess that information and decide what comes next. One can doublethink and give up on eating pizza or can give new commands to the assistant, such as "book a table for two in the restaurant Tasty Pizza, at 7 PM" and "call an Uber to my home at 6:30 PM". It was the human user who chose the restaurant, the time of reservation, the number of seats, and how they preferred to go (cycling, by metro, driving their car, taking an Uber, etc.). For starters, it was the user who decided whether to have a meal that night. A task-automation auxiliary system would not say "eat a salad instead of pizza since you have put on some weight". They work passively, by answering users' questions and performing tasks they are told to. Even when they seem to "guess" what the user wants, it is still based on processing previous parameters and analyzing routine.

Advisory systems are located one step above in the automation chain. They not only help automating tasks but also directly *recommend* what the human user should do, therefore replacing *part* of the human decision-making process. Medical diagnosis software based on imaging tests is a good example. By assessing these images and cross-referring them with other databases, such as scientific literature on recommended treatments for each stage of the illness, the system does not only provide useful information to the medical team but also defines the optimal treatment for each patient. It goes a step further when compared to task-automation auxiliary systems because it takes part in the human decision-making process, automating a substantial part of it. However, here the human still has the final word. Indeed, it is up to the medical team to assess the recommended treatment, explain it to the patient and carry it out. Or even chose a different kind of treatment, by providing scientific base to move away from the system's advice. In common with task-automation auxiliary systems is the fact that advisory systems also work passively (it is up to the medical team and the patient to run it or not). Both are also based on previously fed data instead of real-time sources.

On top of the automation chain are the *full decision-making systems*. They thoroughly replace all or almost all phases of the human decision-making process. It is the AI system and not the human user which makes the calls and carries them out. This kind of system usually *functions actively* and is connected to *real-time data*. Self-driving cars are a perfect example. On the

Some of these assistants, like Google Assistant (which contains Google Lens) and Samsung Bixby, also have the added ability to do image processing to recognise objects in the image to help the users get better results from the clicked images».

highest NHTSA automation level, the vehicle carries out all the decisions in real-time. The human user is merely a spectator with almost no influence in the system's decisions. In some specific contexts, the system will prompt options for the user to choose among them. For instance, to choose between faster paths (less traffic) or safer although longer paths (avoiding dangerous neighborhoods), or if the user wants to change the route in case an unforeseen accident causes a traffic jam. Note that after the user has chosen one of the options it is the self-driving car that does all the rest. This is quite different from task-automation auxiliary systems since they provide information based on what the human user had asked them to do, while in self-driving cars it is the AI system, and not the user, which defines when, where, and which options will be prompted for human choice, based on real-time data. It is also different from advisory systems since *all or almost all* (and not only some) substantial phases of the human decision-making process are replaced by AI.

LEVELS OF AI SYSTEMS INTERFERENCE IN HUMAN DECISION-MAKING

Kind of AI system	**Automation Level**	**Functioning**	**Input**	**Output**	**Interference in human decision-making**
Task-automation auxiliary	Initial	Passive	Past data	Provide information or run human commands	Do not replace any stage of human decision
Advisory	Intermediary	Passive	Past data	Recommend an action	Replace one or more substantial stages of human decision
Full decision-making	Advanced	Active	Real-time data	Makes and executes decisions autonomously	Replaces all or almost all human decision

Figure 1. Levels of AI systems interference on human decision-making (author's creation).

As already mentioned in Section 3, there is currently an expressed or implied widespread understanding that AI-based systems should always surpass human capabilities, no matter the context. In other words, their accuracy rates must beat human standards. That false premise does not account for the practical differences in the 3-level categorization of AI interference in the human decision-making process, as present in this section. It also does not account for the fact that *each purpose* for using an AI system should target not only different *accuracy rates* but also *acceptable risks and transparency, on a case-by-case basis*. It is the reason why I call it a *false* premise. The following sections will dig deeper into that.

5. – *"All models are wrong, but some are useful (…)"*[36].

With this catching phrase from 1979, George Edward Pelham Box highlighted that there is no such thing as a perfect system. Even cutting-edge AI will have some level of inaccuracy, inherent risks, and lack of transparency. Systems that run continuously will sooner or later experience at least some slightest failure, due to internal or external factors. Therefore, 100% accuracy is not feasible. Acknowledging that prevents misalignment between the results people *expect* from AI and what this technology *can actually deliver*. Moreover, it helps demonstrate that there are many contexts in which AI should not beat human standards, since it can be quite useful just by replacing human labor, even at the cost of a substantial decrease in accuracy, risk prevention, or transparency.

Therefore, the key point to assess the suitability of any AI system is to run a *case-by-case analysis* to determine which is the acceptable accuracy, risk prevention, and transparency for each system, considering the context and purpose of use of that system, as well as the AI interference in the human decision-making, according to the author 3-level categorization developed in Section 4. *The following subsections will provide the author's criteria to set these limits*, as well as situations in which mandatory human intervention ("human in the loop" - HITL) should be considered. Bear in mind that the optimum result comes from a mix of these factors, considering the specifics of each context. There is no one-size-fits-all answer.

[36] G.E.P. BOX, *Robustness in the strategy of scientific model building*, in *University of Wisconsin-Madison Mathematics Research Center*, 1999, available at: <https://apps.dtic.mil/sti/citations/ADA070213>. Access: Mar. 10. 2022, 02.

5.1. – As mentioned in Section 3, expecting that AI systems should always reach high accuracy rates is a false premise. Accuracy is just one of the factors to be accounted for when assessing the suitability of a system. *Each context and purpose for using an AI demands a different accuracy rate*. Consequently, it is not desirable nor in accordance with the current level of technological development to expect that any AI-based system reaches accuracy rates of 90% or higher, regardless of the context or purpose of using the system. On the contrary, the analysis should be run on a *case-by-case basis* to determine which is the acceptable accuracy for a given case. The bare minimum in some situations may be the optimum level in others.

For instance, think of a hypothetical face recognition system[37] used to control people's access to a large city (blocking entrance or leaving). Consider that system running with an accuracy rate of 90%[38] in the city of São Paulo, in Brazil, with around 12 million residents. It means that this system will fail and wrongly block the massive amount of 1,200,000 people. Now think about the same system running in a city like Beijing, in China, with the population two times bigger. In those examples, even high accuracy rates such as 90% or 95% could *not* be enough and *should not be allowed* since a minimum lack of accuracy could lead to catastrophic results. The same line of reasoning also applies to real life situations, such as law enforcement agents using software for facial recognition of criminals. In these and many other contexts, only absurdly high accuracy rates such as 98% or 99% are to be expected.

On the contrary, cases are in which accuracy rates of 50% or even less should suffice. For instance, in high inherent risk or unhealthy activities, the replacement of human labor by an AI is justifiable even at the expense of a substantial decrease in accuracy, because it prevents physical or psychological harm. Therefore, even an AI less precise than humans could be quite useful in some contexts. Moreover, a suboptimal AI system coupled with human intervention may lead to excellent results, with final accuracy rates even higher than the level provided by using only the system. This *man*

[37] M. O'Flaherty, *Facial Recognition Technology and Fundamental Rights*, in *European Data Protection Law Review*, Berlin, v. 06, n. 02, 2020, 170: «Facial recognition technology (FRT) makes it possible to compare digital facial images to determine whether they are of the same person. Comparing footage obtained from video cameras (CCTV) with images in databases is referred to as 'live facial recognition technology».

See also: A.K. Jain – A.A. Ross – K. Nandakumar, *Introduction to Biometrics*, New York, 2021.

[38] *Op. cit.*, 172: «Facial recognition technology algorithms *never provide a definitive result*, but only probabilities that two faces appertain to the same person».

+ *machine*[39] situation is another example of acceptable lower accuracy rates for an AI. As it is well known, chess-playing software beat the world champion Kasparov[40]. However, it is less known that the best human chess players using an ordinary computer to assist them in simulating the movements can beat the AI[41]. Thus, *man + machine in some contexts can conquer more than either of them would be able to achieve alone*[42].

[39] S. CAO, *et al*, *From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses*, in *Columbia Business School Research Paper*, 2021, 2: «The existing literature has been mostly focusing on characterizing the type of jobs that are vulnerable to disruption by AI's evolution, as well as those it could create. In other words, the sentiment of the existent studies mostly involves a theme of 'Man versus Machine', which characterizes the contest between humans and AI, explores ways humans adapt, and predicts the resulting job redeployments. In such settings, human beings are often rendered passive or reactive-dealing with disruptions and looking for new opportunities defined by the AI landscape. There has been relatively little research devoted to prescribing how skilled human workers could tap into a higher potential with enhancement from AI technology, which is presumably the primary goal for humans to design and develop AI in the first place. This study aims to connect the contest of 'Man versus Machine' ('Man vs. Machine' hereafter) to a potential equilibrium of 'Man plus Machine' ('Man + Machine' hereafter)».

See also: Y.N. HARARI, *Homo Deus*: *A Brief History of Tomorrow*, New York, 2016, 44: «The upgrading of humans into gods may follow any of three paths: biological engineering, cyborg engineering and the engineering of non-organic beings».

[40] And many other professional gamers… More recently, AI has beaten 8 world champions at bridge: L. SPINNEY, *Artificial intelligence beats eight world champions at bridge*, in *The Guardian*, London, 2022, available at: <https://www.theguardian.com/technology/2022/mar/29/artificial-intelligence-beats-eight-world-champions-at-bridge>. Access: 31 Mar. 2022.

[41] E. BRYNJOLFSSON – A. MCAFEE, *The Second Machine Age*: *Work, Progress, and Prosperity in a Time of Brilliant Technologies*, New York, 2016, 188 ss.

[42] W. BARFIELD, *Cyber-Humans*: *Our Future with Machines*, New York, 2015, 1: «(…) our future is to merge with artificially intelligent machines! How I reached that conclusion is the subject of this book. I don't mean to imply that in the coming decades we humans will look and act like robots on an assembly line, rather, that we will be equipped with so much technology, including computing devices implanted within the brain itself, that we will have been transformed from a biological being into a technology-based being, evolving under laws of technology, more so than under the laws of biological evolution».

See more about mandatory human intervention (human in the loop) in Section 5.3.

Also have in mind that man + machine is a controversial issue: A. FÜGENER, *et al*, *Will humans-in-the-loop become borgs? Merits and pitfalls of working with AI*, in *Management Information Systems Quarterly*, Minnesota, v. 45, n. 03, 2021, 1527: «We analyze how advice from an AI affects complementarities between humans and AI, in particular what humans know that an AI does not know: 'unique human knowledge'. (…) Simulation results based on our experimental data suggest that groups of humans interacting with AI are far less effective as compared to human groups without AI assistance».

In the above-mentioned contexts, putting an excessive focus on accuracy can lead to *unwanted outcomes* such as: hindering the entrance of new players in the market by demanding accuracy rates much higher than their products or services can (and should) provide, unnecessarily stretching the development and testing cycle, rising productions costs and, in many cases, even preventing the deployment of products and services that could reduce human exposure to high inherent risk or unhealthy activities. Summing up, an exaggerated focus on accuracy can compromise innovation, and curtail competitiveness and wellbeing, as recognized be the OECD[43].

Having clarified this point, the next question is: how to define the acceptable accuracy of an AI system, in each context? *The answer comes from weighing two factors*: 1) the level of AI interference in the human decision-making process (as described in Section 4); and 2) the inherent risks of the activity to be automated.

On the one hand, the higher the level of AI interference in the human decision-making process, the *higher* tends to be the need to assure that the system's accuracy surpasses human standards. On the other hand, the higher the inherent risks of the activity to be automated, the *lower* tends to be the system's acceptable accuracy. In other words, *there is an inverse relationship between these factors*. AI interference in human decision-making points to higher accuracy rates, whereas higher inherent risks may justify lower accuracy. That's why it is so crucial to *weigh* these two factors, on a case-by-case basis.

Concerning the first factor, the 3 different levels of AI interference in the human decision-making process were already described in Section 4. In general, higher levels of AI interference demand higher accuracy. On level 1 (task-automation auxiliary systems) *lower accuracy rates are acceptable* since it is the human user (and not the AI) who will assess all the information and make the decision. The system works only as a tool to search for that information and eventually automating some tasks, based on human commands. Differently, on level 2 (advisory systems) at least one substantial part of human decision-making will be completely replaced by the system's output. Therefore, an accuracy higher than in level 1 is expected. Indeed, the precision here should be *at least equivalent to human standards* since a wrong output can compromise the next steps. Finally, on level 3 (full decision-making systems) all or almost all substantial phases of the human decision-

[43] OECD – Organization for Economic Co-Operation and Development. *OECD Framework for the Classification of AI Systems*. Available at: <https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en>. Access: 20 Jun. 2022, 67: «Policy makers favour a risk-based approach to regulating AI in order to focus oversight and intervention where it is most needed, while avoiding unnecessary hurdles to innovation».

making process are replaced by AI. In this context, it is logical to expect that the system's output has an accuracy rate *higher than human standards*. After all, it would not make sense to replace human labor with a system that performs worse, since automation should seek efficiency.

Concerning the second factor, each activity has a level of inherent risks[44], which means risks naturally associated with that activity that can eventually be mitigated but not 100% prevented[45]. When the level of inherent risks is high it makes sense to accept lower (or much lower) accuracy rates from an AI system when compared to human standards. In this context, gains in other factors, such as risk prevention or transparency can compensate for the loss of precision. After all, preserving the physical and psychological integrity of the human being is a goal that should prevail even at the expense of accuracy decrease. A classic example is a robot designed to defuse bombs. Even if its accuracy is lower than human standards, it is still justifiable to use the robot instead of a human expert to reduce the risk of injuries or death to that expert. Several other inherent risks or unhealthy activities fit into this line of reasoning.

Therefore, it is vital to weigh *both* the level of AI interference in the human decision-making process and the level of inherent risks of an activity to set which is the acceptable accuracy of an AI in each context[46]. In some cases, risk reduction may justify accuracy rates *lower* than human standards, as well as on levels 2 and 3 of automation, such as in the example of a full decision-making robot to defuse bombs. In face of a high risk that is hardly mitigable, it seems a better option to expose the corpus of a robot to that risk instead of a human being[47], even at the expense of a substantial decrease in accuracy. In other cases, it suffices that the system has accuracy *equivalent* to humans, therefore generating free time without a substantial decrease in efficiency. An example is a software used in courts to group lawsuits according to the legal issue discussed, time of filling, or any other parameter.

[44] ID., «The risks in using any AI system strongly depend on the application. Since it is difficult to anticipate and assess every possible use case, applied AI systems should be grouped into a small collection of risk levels».

[45] T.A. LOPEZ, *Princípio da Precaução e Evolução da Responsabilidade Civil*, São Paulo, 2010, 25: «Risk is the eventual danger that is more or less predictable, and it is different from "álea" (unpredictable) and from the danger (actual). The risk is abstract»; C. CATH, *et al*, *Artificial Intelligence and the 'Good Society': the US, EU, and UK approach*, in *Science and Engineering Ethics*, New York, v. 23, n. 02, 2017, 21: «AI can easily become the elephant in the crystal room, if we do not pay attention to its development and application».

[46] This understanding is in line with international regulatory initiatives, such as the EU´s proposal for an AI framework. As well as the Brazilian Strategy for Artificial Intelligence. Both will be briefly described in Section 6.

[47] Kate Darling and other social robot enthusiasts may disagree.

To be useful, this kind of software only needs to achieve accuracy rates similar to the employee whose job the AI will replace, therefore freeing that employee to devote more time to other activities, presumably becoming more productive. Finally, in cases where the inherent risk of the activity to be automated is high and failure can compromise fundamental rights, such as in medical imaging tests or robotic surgery, the *maximum accuracy rate available* should be mandatory. In other words, ethical and legal frameworks should forbid human exposure to AI systems when there is scientific evidence that humans performing the same task can achieve better results, without compromising any of the stakeholders (what renders this example different from the bomb-defusing robot). Especially on level 3 of automation, since it is the system, and not the user, which assesses real-time data and makes the decision. This line of reasoning could prevent catastrophic situations, such as in the case of Mracek versus Bryn Mawr Hospital[48].

5.2. – Although the notions of *transparency*[49] and *explainability*[50] are technically different[51], this section addresses them altogether since they are

[48] THE UNITED STATES OF AMERICA. Mracek *v.* Bryn Mawr Hosp. *United States Court of Appeals for the 3rd Circuit*. 610 F. Supp. 2d 401, j. 28.01.2010. Available at: <https://casetext.com/case/mracek-v-bryn-mawr-hosp>. Access: Mar. 19. 2022.

In this well-known case from 2010, Roland C. Mracek was submitted to surgery in the Bryn Mawr Hospital, using the so-called "da Vinci surgical robot". He claimed to be awake at the beginning of the surgery and have seen the robot displaying error messages. He adds that the medical team tried to reboot the robot, but the error messages remained prompting on the screen. They also placed a call to tech support and a representative of the robot's manufacturer came to the operating room but was unable to solve the problem. As a result of the machine's malfunction, and after around 45 minutes, the surgical team abandoned its attempt at a robotic surgery and did it manually. The outcome was tragic: Mr. Mracek suffered permanent damage and had to live with daily pain. He then decided to file a lawsuit against the hospital, based on strict products liability and negligence, claiming that the malfunctioning of the robot was crucial to cause the damage. In the end, the court granted the defendant's motion since Mr. Mracek did not present evidence of a causal relationship between the robot's failure and the results of his surgery.

[49] *Op. cit.*, 83: «A model is considered to be transparent if by itself it is understandable (…)».

[50] ID.: «Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans».

D. LEHR – P. OHM, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning, University of California Davis Law Review*, Davis, v. 51, n. 02, 2017, 705 ss.: «(…) "explainability" [is the] the ability of machine learning to give reasons for its estimations».

intimately connected. Roughly speaking, they mean that a human user is able to understand why an AI system generated a certain output and explain it to an ordinary user of that system. Transparency is undoubtedly a fundamental value provided in numerous legal standards, for both the public[52] and the private[53] sectors, worldwide. It is of paramount importance and should be respected according to the provisions of each legal system. Therefore, the bigger the transparency, the better[54].

However, cases are in which the AI system still *cannot* provide high transparency rates (such as 90% or higher), due to the current level of technological development. This is a fact. But some systems can be quite useful (and may be used) despite their transparency shortcomings. In the bomb-defusing robot example, the value of preventing serious injury to a human being may compensate for the lack of transparency, especially if the robot's accuracy is satisfactory. Many other real-life situations with high inherent risk follow that reasoning.

Thus, transparency is a core factor when assessing an AI's suitability, as well as accuracy. None of them, however, is an end in itself. Consequently, the debate is about which transparency rate should be reasonably expected from AI in each context. To properly answer that question, it is necessary to mention the concept of the "black box" and the trade-off between accuracy and transparency.

[51] UNESCO – UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION. *UNESCO Recommendation on the Ethics of Artificial Intelligence*. Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000380455>. Access: 19 Dec. 2021., 10: «Transparency means allowing people to understand how AI systems are researched, designed, developed, deployed, and used, appropriate to the use context and sensitivity of the AI system. It may also include insight into factors that impact a specific prediction or decision, but it does not usually include sharing specific code or datasets. In this sense, transparency is a socio-technical issue, with the aim of gaining trust from humans for AI systems.

Explainability refers to making intelligible and providing insight into the outcome of AI systems. The explainability of AI models also refers to the understandability of the input, output and behaviour of each algorithmic building block and how it contributes to the outcome of the models. Thus, explainability is closely related to transparency, as outcomes and sub processes leading to outcomes should be understandable and traceable, appropriate to the use context».

[52] For instance, in Brazil the Federal Constitution of 1988 and the Access to Information Act n. 12,527/2011 impose a duty of transparency to the public administration.

[53] Transparency is also a core topic in consumer laws, privacy/data protection laws, labor laws, AI regulation and many other fields, in numerous countries.

[54] Of course, each legal system can legitimately restrict transparency to safeguard other fundamental values, such as industrial and trade secrets. These are exceptions, though.

There is quite a debate in the scientific literature about the fact that some AI-based systems, depending on how they were designed, may become a *"black box"*[55], meaning that it would be hard if not impossible for a human being – even for the developers of the system – to understand the exact reason why it generated a given output[56]. This could be extremely harmful to those system users and the society, for instance, if the system at stake increases discrimination[57] or any other unlawful results, be they intentional or not[58]. Therefore, AI's "black box" is a matter of great concern.

[55] The very concept of "black box" is controverse. There is a general and worldwide known definition: F. PASQUALE, *The Black Box Society*: *The Secret Algorithms That Control Money and Information*, Cambridge, 2015, 3: «The term "black box" is a useful metaphor for doing so, given its own dual meaning. It can refer to a recording device, like the data- monitoring systems in planes, trains, and cars. Or it can mean a system whose workings are mysterious; we can observe its inputs and outputs, but we cannot tell how one becomes the other. We face these two meanings daily: tracked ever more closely by firms and government, we have no clear idea of just how far much of this information can travel, how it is used, or its consequences».

There are subclasses of black box: G.N. LA DIEGA, *Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information*, in *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, Göttingen, v. 09, n. 01, 2018, 9: «The lack of transparency is related to the so-called black box (better said, black boxes). Arguably, three different black boxes may be distinguished: the organisational; the technical; and the legal one».

And there is also the concept of black box related to proprietary content, such as trade secrets: C. RUDIN, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, in *Nature Machine Intelligence*, 2019, 206: «A black box model could be either (1) a function that is too complicated for any human to comprehend or (2) a function that is proprietary».

[56] J. BLACK – A. MURRAY, *Regulating AI and Machine Learning: Setting the Regulatory Agenda*, in *European Journal of Law and Technology*, Coventry, v. 10, n. 03, 2019, 7 ss.: «One clear systemic risk of AI and ML [machine learning] is the "black box" issue. This is the problem that arises when an algorithmic system makes decisions which prove extremely difficult to explain in a way that the average person can understand. In essence while it is possible to observe incoming data (input) and outgoing data (output) in algorithmic systems, but their internal operations are not very well understood».

[57] Some authors point out that it would be theoretically easier to prevent discrimination using an AI than to prevent it in human behavior, since the AI could be designed from the beginning to be auditable: C.R. SUSTEIN, *Discrimination in the Age of Algorithms*, in *Journal of Legal Analysis*, Cambridge, v. 10, n. 01, 2018, 113 ss.: «Our central claim here is that when algorithms are involved, proving discrimination will be easier – or at least it should be, and can be made to be. The law forbids discrimination by algorithm, and that prohibition can be implemented by regulating the process through which algorithms are designed. This implementation could codify the most common approach to building machine-learning classification algorithms in practice, and add detailed record-keeping requirements. Such an

When analyzing the alternatives to solve that problem, some authors point out that there would be an inevitable *trade-off between accuracy and transparency*. They advocate that higher accuracy rates tend to produce black boxes, while fully explainable systems would be less accurate[59]. In other words, enhancing one of them would decrease the other. Differently, other authors indicate that a focus on transparency as a core goal during *all phases* of product development (*transparency by design*)[60] can assure higher accuracy

approach would provide valuable transparency about the decisions and choices made in building algorithms – and also about the tradeoffs among relevant values. (…) Getting the proper regulatory system in place does not simply limit the possibility of discrimination from algorithms; it has the potential to turn algorithms into a powerful counterweight to human discrimination and a positive force for social good of multiple kinds».

[58] Y. BENKLER, *Don't let industry write the rules for AI*, in *Nature*, v. 569, 2019, 1: «Inside an algorithmic black box, societal biases are rendered invisible and unaccountable. When designed for profit-making alone, algorithms necessarily diverge from the public interest – information asymmetries, bargaining power and externalities pervade these markets».

[59] L. EDWARDS – M. VEALE, *Slave to the Algorithm? Why a 'Right to an Explanation' is probably not the remedy you are looking for*, in *Duke Law & Technology Review*, Durhamv. 16, n. 01, 2017, 59 ss.: «Meaningful explanations of ML [machine learning] do not work well for every task. (…) the tasks they work well on often have only a few input variables that are combined in relatively straightforward ways, such as increasing or decreasing relationships. Systems with more variables will typically perform better than simpler systems, so we may end up with a trade-off between performance and explicability (…) Optimising an explanation system for human interpretability necessarily means diluting predictive performance to capture only the main logics of a system».

V. ALMEIDA, *et al*, *A framework for benchmarking discrimination-aware models in machine learning*, in *Artificial Intelligence, Ethics, and Society Conference*, 2019: «It is also expected that the accuracy will be higher than the resulting accuracy when a discrimination-aware technique is used because the technique must lower the discrimination at the cost of accuracy».

[60] D. LEHR – P. OHM, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, in *University of California Davis Law Review*, Davis, v. 51, n. 02, 2017, 668: «With such a singular focus on the running model and a failure to consider stages of machine learning after data management, scholars have been forced to adopt an overly narrow view of algorithms' potential harms and benefits. Inaccuracy and bias are paid much attention, and they can indeed be traced back in part to poor data and variable specifications. But they can also creep in during other stages of machine learning, and many harms arise almost entirely during those other stages. In fact, some of the most viscerally unsettling harms of machine learning – its opacity and lack of explainability – are brought about when algorithms are chosen and developed, not when data are collected or variables are specified».

C. RUDIN, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, in *Nature Machine Intelligence*, v. 1, 2019, 206 ss.: «I am concerned that the field of without curtailing transparency[61]. Therefore, they state that the alleged trade-off is an excuse to justify opaque systems. They add that most (if not all) AI systems become opaque because developers did not pay proper attention to transparency during the development cycle. Consequently, to this line of reasoning black boxes are the result of a development failure.

Having contextualized the concept of the "black box" and the alleged tradeoff between accuracy and transparency, the next step is to present the author's opinion about what should we reasonably expect from AI in terms of transparency. Again, the level of AI interference in the human decision-making process is a key factor that should be considered, on a case-by-case basis. Indeed, AI with *low* interference in the human decision-making process tends to *tolerate less transparency* without compromising the system's suitability. Conversely, the bigger the system interference the bigger the expected transparency. Let's develop this idea a little bit.

On the one hand, it is not reasonable to expect top-notch transparency from systems at level 1 of automation[62] such as Siri and Alexa voice assistants[63], for at least two reasons. First, this kind of system provides less

interpretability/explainability/comprehensibility/transparency in ML [machine learning] has strayed away from the needs of real problems. (…) Recent work on the explainability of black boxes – rather than the interpretability of models – contains and perpetuates critical misconceptions that have generally gone unnoticed, but that can have a lasting negative impact on the widespread use of ML models in society. (…) An inaccurate (low-fidelity) explanation model limits trust in the explanation, and by extension, trust in the black box that it is trying to explain».

[61] ID.: «It is a myth that there is necessarily a trade-off between accuracy and interpretability. There is a widespread belief that more complex models are more accurate, meaning that a complicated black box is necessary for top predictive performance. However, this is often not true, particularly when the data are structured, with a good representation in terms of naturally meaningful features. When considering problems that have structured data with meaningful features, there is often no significant difference in performance between more complex classifiers (deep neural networks, boosted decision trees, random forests) and much simpler classifiers (logistic regression, decision lists) after preprocessing».

[62] As described in Section 4.

[63] The scientific literature usually also quotes the example of map applications, such as Google Maps: C. O'NEIL, *Weapons of Math Destruction*, New York, 2016, 3 ss.: «There would always be mistakes, however, because models are, by their very nature, simplifications. No model can include all of the real world's complexity or the nuance of human communication. Inevitably, some important information gets left out. (…) To create a model, then, we make choices about what's important enough to include, simplifying the world into a toy version that can be easily understood and from which we can infer important facts and actions. We expect it to handle only one job and accept that it will occasionally act like a clueless machine, one with enormous blind spots. (…) Sometimes these blind spots don't matter. When we ask Google Maps for directions, it models the world as a series of roads, tunnels, and bridges. It ignores the buildings, because they aren't relevant to the task».

technical day-to-day information. It is the human user who will assess that information and make the decision. In case of an unreasonable output from the system, the user can cross-reference that information with other easy-to-find sources and make a critical assessment before deciding[64]. Second, it is expected that these systems are involved in commercial agreements entered by the developer. Accordingly, it shouldn't be surprising if an Apple voice assistant favored advertising information related to Apple's business partners over similar information related to competitors. The same for Google, Amazon, or any other major market player. As long as there is no unlawful competition in place[65], these commercial arrangements are legitimate.

On the other hand, in AI that plays a significant role in replacing human decision making, such as in levels 2 and 3 of automation, lack of transparency can be an issue even if the system accuracy is high. Opaqueness by itself can jeopardize protected values, since affected people have *the right to know*[66] why the system made a certain decision, especially if

[64] It is known that the real world is so complex that cases will arise in which a wrong output from a vocal assistant can be devastating. These cases, though, are exceptions.

[65] For instance, a joint report from University of Washington, UC Davis, UC Irvine and Northeastern University in 2022 pointed out that Alexa data was unlawfully shared with Amazon commercial partners, against US privacy laws. See: J.P. TUOHY, *Researchers find Amazon uses Alexa voice data to target you with ads*, in *The Verge*, 2022, available at: <https://www.theverge.com/2022/4/28/23047026/amazon-alexa-voice-data-targeted-ads-research-report>. Access May. 30, 2022.

[66] In privacy/personal data protection regulations around the world, this provision is usually called "a right to revision of automated decision-making" or "a right to explanation". For instance, in the European General Data Protection Regulation of 2016 (GDPR - Regulation 2016/679) article 22; in the Brazilian General Data Protection Act of 2018 (LGPD - Federal Law n. 13,709/2018) article 20; and in Ley n. 18,331 of 2008 from Uruguay article 16, among many others.

The aforementioned provision is highly controversial. On the one hand, some authors point out that it does not grant the interested party a right to know exactly how the algorithm works, since it would conflict with intellectual property rights of the algorithm's owner: S. WACHTER – B. MITTELSTADT – C. RUSSELL, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, In *Harvard Journal of Law & Technology*, Cambridge, v. 31, n. 02, 2018, 863 ss.: «The description of explanations in Recital 71 does not include a requirement to open the 'black box'. Understanding the internal logic of the algorithmic decision-making system is not explicitly required».

On the other hand, some authors advocate for a broader right to explanation, especially when concerning the public administration: M. FERRARI, *L'uso degli algoritmi nella attività amministrativa discrezionale*, in *Questa Rivista*, 2020, 1, 67: «(...) è determinante che ciascuno dei soggetti coinvolti, sia la PA [Pública Amministrazione] che i singoli/privati coinvolti dalla procedura, siano in grado di conoscere il meccanismo di operatività dell'algoritmo, anche se per motivazioni they suspect that some unlawful parameter was used, whether the decision was accurate or not. Consider, for instance, a medical imaging software that recommends the optimal treatment for each patient. Even if the medical team agrees with the software recommendation, they still should be able to *explain* to the patient why the system made that choice. Moreover, they should be able to spot inconsistencies such as bias or unreasonable results, therefore suggesting a correction.

Summing up, *lower replacement* of human decision-making process tends to *tolerate less transparency* since it is easier for the users to prevent harm by reviewing other sources of information before making up their minds and it is expected that this kind of system is lawfully engaged in commercial agreements entered by the developer. Quite the opposite happens with systems that play a *significant role in replacing human decision-making*, such as in levels 2 and 3 of automation, since transparency shortcomings may be illegal in themselves, irrespective of the system's accuracy. Moreover, users should be fully informed about commercial agreements related to these systems and these agreements must respect strict ethical and legal boundaries. Therefore, in this second scenario, *the maximum transparency rate available should be mandatory*.

Finalizing this section, it is worth mentioning that regulation should demand compliance with minimum transparency standards during all the development cycle, for any AI system, to facilitate risk assessment and foster transparency by design. This is quite a consensus in the legal field. The author's point is that those standards *don´t* need to be the same for all kinds of systems. On the contrary, each system calls for a different transparency rate, on a case-by-case basis. Indeed, factors such as the many ways to provide AI-based products and services into the market, the purpose of using each system, the different levels of AI interference in human decision-making, accuracy rate, inherent risks of the activity to be automated, and transparency are *components of a formula* to assess the suitability of an AI, by balancing the interests and expectations at stake. In this formula, transparency is undoubtedly a key factor. It is just not the only one.

5.3. – Even when an AI-based system reaches high accuracy and transparency rates, such as in the examples mentioned in Sections 5.1 and 5.2, there are still contexts in which legal or social reasons may impose –

diverse fra loro: per la PA è necessario comprendere se quell'algoritmo consenta di centrare, legittimamente con equità, gli obiettivi che ci si proponeva di raggiungere con la procedura amministrativa avviata; per il singolo cittadino è utile sapere come sia stata processata la scelta amministrativa, per poter escludere di essere stati vittime di ingiuste incongruenze con conseguenziali lesioni di diritti fondamentali».

through regulation[67] – a prohibition to use that system or, at least, require it to be developed in a way that ensures a meaningful human intervention, if necessary, to override the system's decision[68]. The *total prohibition* of using an AI, in a given context, was called a *ban on AI autonomy*. Diversely, the alternative in which using the system is allowed under the condition that there is room for a *meaningful human intervention* was labeled *human in the loop* - HITL.

Ban on AI autonomy is a radical measure admissible only when the use of an AI-based system is *intrinsically incompatible* with fundamental values, meaning that the system, by its nature, conflicts with human rights, irrespective of the purpose for using it. The classic example is a *Lethal Autonomous Weapon – LAW*[69], a level 3 system[70] for military purposes, such as a drone, a missile, a vehicle, or another kind of embodied AI[71] that carries a weapon, and once activated it is the system itself that takes on searching for the target and engaging in an attack, eventually causing serious injuries or

[67] C. Kopp – M. Lodge, *What is regulation? An interdisciplinary concept analysis*, in *Regulation & Governance*, Hoboken, v. 11, n. 01, 2015, 20 ss.: «*(…)* we can distinguish two types of definitions that cut across disciplines – an *essence-based* and a *pattern-based* definition of regulation. (…) an essence-based definition aims to capture the minimal essence of the concept. It is a classical definition in the sense that it includes – and solely includes – those elements without which regulation loses its identity (…). Accordingly, regulation can be defined as *the intentional intervention in the activities of a target population*. The intervention which this definition refers to can be direct and/or indirect, the activities can be economic and/or non-economic, the regulator may be a public or private-sector actor, and the regulatee may equally be a public or private-sector actor. (…) Our pattern-based definition is not less inclusive than the essence-based one, but it gives insight into the manifestations which regulation scholars are mainly concerned with, and which we consider more central to the concept. Attributing more importance to the variation in emphasis of studies, regulation can be defined as *the intentional intervention in the activities of a target population, where the intervention is typically direct – involving binding standard-setting, monitoring and sanctioning – and exercised by public-sector actors on the economic activities of private-sector actors*».

See, also: L. Parentoni, *Artificial Intelligence*, in *Encyclopedia of the Philosophy of Law and Social Philosophy*, Dordrecht, 2020.

[68] M. Wimmer – D. Doneda, *"Falhas de IA" e a Intervenção Humana em Decisões Automatizadas: Parâmetros para a Legitimação pela Humanização*, in *Revista Direito Público*, Brasília, v. 18, n. 100, 2021, 384. Loosely translated from the original, in Portuguese: «(...) even if a given autonomous system reaches an acceptable level of hit and miss rates, would it be ethically legitimate to delegate certain types of decision entirely to automated systems, without relevant human intervention?».

[69] Also known as Autonomous Weapons System – AWS.

[70] According to the classification provided in Section 4.

[71] According to the definition of embodied AI provided in Section 2.

even death[72]. This is the kind of AI that most resembles the "evil machines" depicted in movies such as Terminator.

Although there is no consensus on the matter in scientific literature[73], the US Department of Defense considers the ability to, once activated, select, track, and engage targets without further human intervention as the core attribute of an autonomous weapon[74]. China proposed a more detailed list, with 5 basic attributes:

> *"1) lethality, which means sufficient pay load (charge) and for means to be lethal;*
>
> *2) autonomy, which means absence of human intervention and control during the entire process of executing a task;*
>
> *3) impossibility for termination, meaning that once started there is no way to terminate the device;*
>
> *4) indiscriminate effect, meaning that the device will execute the task of killing and maiming regardless of conditions, scenarios and targets; and*
>
> *5) evolution, meaning that through interaction with the environment the device can learn autonomously, expand its functions and capabilities in a way exceeding human expectations."*[75]

Irrespective of the definition adopted, the crucial matter here is the fact that a decision about taking a life, even during a war, should be made by

[72] N. Davison, *A legal perspective: Autonomous weapon systems under international humanitarian law*, in *UNODA Occasional Papers n. 30*, 2018, 6: «After initial launch or activation by a human operator, it is the weapon system itself – using its sensors, computer programming (software) and weaponry – that takes on the targeting functions that would otherwise be controlled by humans. This working definition encompasses any weapon system that can independently select and attack targets, including some existing weapons and potential future systems».

[73] S.R. Reeves – R.T.P. Alcala – A. McCarthy, *Challenges in regulating lethal autonomous weapons under international law*, in *Southwestern Journal of International Law*, Los Angeles, v. 27, n. 01, 2021, 105: «The concept of autonomous weapon systems is itself not clearly defined internationally (…)».

J. Cherry – D. Johnson, *Maintaining command and control (C2) of lethal autonomous weapon Systems: Legal and policy considerations*, in *Southwestern Journal of International Law*, Los Angeles, v. 27, n. 01, 2021, 6: «Definitions abound for autonomous weapon systems among the international legal and policy communities, but States have struggled to agree on a common definition».

[74] THE UNITED STATES OF AMERICA. *Department of Defense – DoD Directive 3000.09*. Available at: <https://www.hsdl.org/?abstract&did=726163 >. Access: 04 Jun. 2022.

[75] CHINA. *CCW/GGE.1/2018/WP.7*. Available at: <https://undocs.org/Home/Mobile?FinalSymbol=CCW%2FGGE.1%2F2018%2FWP.7&Language=E&DeviceType=Desktop&LangRequested=False>. Access: 04 Jun. 2022.

humans, and by humans only. For both ethical and legal reasons[76]. Therefore, lethal autonomous weapons are the perfect example of a system that should be subject to a ban on AI autonomy[77]. So much so that specialists in humanitarian law advocate for an international treaty to amend the UN Convention on Conventional Weapons to deal with that matter, worldwide[78]. Civil society organizations such as Human Rights Watch also support that proposal[79].

Other examples of a ban on AI autonomy are even more controversial. One of them is the *court's decision,* meaning the hypothesis in which the human judge could be completely replaced by an AI system that would drive the process and make all the decisions[80]. Lawrence Solum goes one step ahead, proposing the debate around a hypothetical "*artificially intelligent law*"[81], an entire legal system driven by AI. Considering that the judicial

[76] D.K. CITRON – F. PASQUALE, *The Scored Society: Due Process for Automated Predictions, University of Maryland School of Law Research Paper n. 214-8,* 2014, 7: «Human rights advocates and computer scientists contend that 'Human-out-of-the-Loop Weapons' systems violate international law because AI systems cannot adequately incorporate the rules of distinction ("which requires armed forces to distinguish between combatants and noncombatants") and proportionality».

[77] I respectfully disagree with authors who support the use of those systems, such as: D. MACINTOSH, *Fire and Forget: A Moral Defense of the Use of Autonomous Weapons Systems in War and Peace,* in *Lethal Autonomous Weapons*: *Re-Examining the Law and Ethics of Robotic Warfare,* Oxford, 2021, 9: «I want to argue more specifically (…) that there are many conditions where using AWSs would be appropriate not just rationally and strategically, but also morally».

[78] THE UNITED NATIONS. *Background on LAWS in the CCW.* Available at: <https://www.un.org/disarmament/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/>. Access: 04 Jun. 2022.

[79] They published a manifesto called "Stop Killer Robots": HUMAN RIGHTS WATCH. *New Weapons, Proven Precedent - Elements of and Models for a Treaty on Killer Robots.* Available at: <https://www.hrw.org/report/2020/10/20/new-weapons-proven-precedent/elements-and-models-treaty-killer-robots >. Access: 05 Jun. 2022: «A majority of CCW states parties and the Campaign to Stop Killer Robots, a global civil society coalition coordinated by Human Rights Watch, are calling for the negotiation of a legally binding instrument to prohibit or restrict lethal autonomous weapons systems. The Campaign advocates for a treaty to maintain meaningful human control over the use of force and prohibit weapons systems that operate without such control».

[80] P.R.B. FORTES, *et al, Artificial Intelligence Risks and Algorithmic Regulation,* in *European Journal of Risk Regulation,* Cambridge, v. 13, n. 02, 2022, 11: «In this sense, AI would be trained for the specific task of providing judicial decision-making».

[81] L.B. SOLUM, *Artificially Intelligent Law,* in *SSRN Research Paper,* 2019, 53 ss.: «This paper explores a series of thought experiments that postulate the existence of "artificially intelligent law". An artificially-intelligent legal system is defined as one with three functional capacities: 1. The system has the capacity to generate legal norms. 2. The system has the capacity to apply the legal norms that it generates. 3.

process deals with intrinsically subjective judgments, and that case law must be constructed by humans (and not by machines), this is a field in which a ban on AI autonomy makes sense.

A less controversial example presents itself in the health sector, with *autonomous robotic surgeries*, which means a surgery fully conducted by robots, without meaningful human intervention. In this context, a ban on AI autonomy makes sense especially when accuracy rates are low or when there is not enough transparency regarding how the system works[82]. This study will not investigate these and other possible cases of a ban on AI autonomy, since they would demand a deeper analysis, falling outside the scope of the current research.

After briefly contextualizing the ban on AI autonomy, it is time to address its main alternative, known as human in the loop – HITL.

Indeed, nowadays humans take part in the *initial* programming of an algorithm, the construction of the algorithm itself[83]. In this sense, there would always be a human in the loop, since irrespective of the output produced by the AI, it would have been preceded by some level of human intervention. This is *not* the meaning of HITL adopted in this study. As a matter of fact, *HITL is the possibility of a human being monitoring an AI system, being able to suspend it at any time, or even intervene in the system's decisions, to override them.* It means *ensuring a meaningful human intervention* over the system, during or after its functioning[84]. The aspect of *monitoring* is usually called human *on* the loop, while the *intervention* stands for human *in* the loop[85]. Putting aside this technicality, human in the loop is the most used expression, encompassing both meanings. At the end of the day, what

The system has the capacity to use deep learning to modify the legal norms that it generates. (…) Delegating the law-making function to an artificial intelligence is qualitatively different than any current use of artificial intelligence of which I am aware».

[82] As seen in Sections 5.1 and 5.2.

[83] At least for a while, until evolutionary algorithms eventually become able to develop alone.

[84] Although HITL should be implemented by design, to work properly.

[85] J. Cherry – D. Johnson, *Maintaining command and control (C2) of lethal autonomous weapon Systems: Legal and policy considerations*, in *Southwestern Journal of International Law*, Los Angeles, v. 27, n. 01, 2021, 4 ss.: «In supervised autonomous operation, or 'human *on* the loop', the machine can sense, decide, and act on its own once put into operation, but a human user can observe the machine's behavior and intervene to stop the action if necessary. Supervised autonomous robotic surgery is an example of a supervised-autonomous system. In the last degree, fully autonomous operation, the system can sense, decide, and act without human intervention. The human is "*out* of the loop" in that the machine operates without communicating back to the human user. A Roomba vacuum is an example of a fully autonomous system».

stands out is the fact that the final decision is restricted to a human being, whenever necessary.

Instead of just banning a system, it can be used under the condition that HITL is in place. For instance, with *irreversible or hard-to-reverse decisions*, such as hiring or firing an employee, since it requires balancing factors that are both objective (such as the economic situation of the company, worker's rate of absence, and productivity) and subjective (social environment, family and health situation of the employee, etc.). This kind of balance demands "humanity" in the decision. Therefore, HITL is mandatory for both ethical and legal reasons. So much so that the International Labor Organization – ILO Code of Practice on the Protection of Workers Personal Data of 1997 mandates a meaningful human intervention to recruit or dismiss workers[86]. More recently, a Committee at the Council of Europe called member countries to forbid facial recognition systems lacking proper human intervention[87].

The already mentioned *robotic surgery* illustrates how HITL can serve as an alternative to banning AI autonomy. Considering that AI failures here can cause a tragic outcome, such as in the case Mracek versus Bryn Mawr Hospital[88], a meaningful human intervention may be the last resource in a life-or-death situation. Therefore, regulation should impose HITL instead of banning the use of level 3 autonomous surgical robots.

A third example of HITL is the *administrative decision*, thus considered as the decision made by the public administration, such as government agencies while exercising their institutional powers. As it is well known, the constitution of many countries demands that this kind of decision should aim at the public interest and be justified by the public agent. Therefore, even if using an AI to fully replace human decision-making is a viable

[86] THE INTERNATIONAL LABOUR ORGANIZATION. *Code of Practice on the Protection of Workers Personal Data*. Available at: <https://www.ilo.org/global/topics/safety-and-health-at-work/normative-instruments/code-of-practice/WCMS_107797/lang--en/index.htm>. Access: 05 Jun. 2022: «5.5. Automated procedures do not absolve employers from consulting all the data necessary to evaluate correctly the results of the processing. The code thus rejects a purely mechanical decision-making process and opts instead for a clearly individualized evaluation of workers».

[87] EUROPEAN UNION. Council of Europe. *Consultative Committee of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data*. Brussels: 28 Jan. 2021. Available at: <http://rm.coe.int/0900001680a134f3>. Access: 05 Jun. 2022., 5: «The level of intrusiveness of facial recognition, and related infringement on the rights to privacy and data protection will vary according to the particular situation of their uses and there will be cases where domestic law will strictly limit it, or even completely prohibit it where the democratic process will have led to that decision».

[88] Described in Section 5.1.

option, it remains necessary to ensure a meaningful human intervention, to correct system failures capable of causing damage to citizens or to the public administration itself[89].

A fourth example, extremely controversial, is *content moderation in online social networks*, such as Facebook, Instagram, or Twitter. Due to the high amount of data involved, this moderation is already algorithmic-driven, according to the terms of use of each platform. However, in case of a failure - for instance when legitimate content is blocked - how should be the proceeding of human intervention? This is a topic debated worldwide, with a clear political component[90].

Another controversial case involved the Uber transport app, in 2021. To verify if the person trying to log in was the registered driver, the company used a facial recognition system that required sending a selfie, in real-time. However, that system failed at recognizing selfies from people of color. Only after human intervention by an Uber employee were the affected drivers allowed to log in[91].

Lastly, one of the fields in which HITL is already rooted is *personal data protection*[92]. For instance, the European regulation (GDPR) addresses the subject as follows:

"Article 22. Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

[89] For a deeper discussion about AI in the public sector, see: M. FERRARI, *L'uso degli algoritmi nella attività amministrativa discrezionale*, in *Questa Rivista*, 1, 2020.

[90] The Facebook's oversight board is a good study-case. Created by the company in 2020, it gathers renowned international experts, with different backgrounds and cultures. Its mission is to: «[use] its independent judgment to support people's right to free expression and ensure those rights are being adequately respected. The board's decisions to uphold or reverse Facebook's content decisions will be binding, meaning Facebook will have to implement them, unless doing so could violate the law». See more in: https://oversightboard.com/

[91] Some UK unions sued Uber for algorithmic discrimination. See: https://www.business-humanrights.org/en/latest-news/uk-drivers-couriers-sue-uber-over-allegedly-racist-facial-recognition-checks/

[92] M. WIMMER – D. DONEDA, *"Falhas de IA" e a Intervenção Humana em Decisões Automatizadas: Parâmetros para a Legitimação pela Humanização*, in *Revista Direito Público*, Brasília, v. 18, n. 100, 2021, 385. Loosely translated from the original, in Portuguese: «(..) it is in the field of personal data protection that attempts to promote the introduction of "human" elements in decisions taken automatically can be observed more clearly».

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

(c) is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision."[93]

A similar provision can also be found in Directive nº 2016/680, concerning "the protection of natural persons with regard to (…) the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties"[94]. Among other regulations.

Having contextualized human in the loop, it is time to comment on its tradeoff. On the one hand, some authors point out that *man + machine* is of paramount importance to achieve the most of AI[95], such as in the examples previously mentioned. In all of them, assuring a meaningful human intervention over the system is crucial to find a balance between the use of AI and respecting core values protected by law. On the other hand, some

[93] EUROPEAN UNION. European Parliament. *Regulation nº 2016/679*. Brussels: 27 Apr. 2016. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1528874672298&uri=CELEX%3A32016R0679>. Access: 13 Jul. 2018.

In Brazil, the Data Protection Act – "Lei 13,709/2018 (LGPD)" contains a similar provision in article 20. However, during the legislative process Congress has removed the expression that demanded "revision by a natural person", opening room for a debate on the possibility of a software reviewing an automated decision, without any human in the loop.

[94] EUROPEAN UNION. European Parliament. *Directive nº 2016/680*. Brussels: 27 Apr. 2016. Available at: <https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A32016L0680>. Access: 13 Jul. 2018. Article 11.

[95] M.L. AMBROSE, *Regulating the Loop*: *Ironies of Automation Law*, in *WeRobot*, 2014, 16 ss.: «(…) there are extraordinary gains to be made by creating human-machine teams, as the open chess tournament victors exemplify. In order to reach optimization, human-machine systems should be understood as socio-technical systems that do not ignore the human or social contribution to automation and vice versa. (…) Similar to the way humans work with other humans (supported by automated and non-automated tools) to perform the numerous tasks and achieve the many goals in their daily lives, humans will work with robotic and intelligent systems to go about their daily lives – creating a loop in which they are necessarily a part. These loop actors are intertwined».

warn that the high volume of data at stake and the speed of processing surpass the human brain's capabilities, which renders human intervention slow, ineffective, or even impossible[96]. They advocate that the aim of automation through AI has been achieving faster and better results. Therefore, they sustain that HITL would be a contradiction since it slows down the system and curtails efficiency[97]. In fact, both lines of reasoning are partially correct. *HITL undoubtedly comes at a cost*. It is an *alternative* to avoid banning AI autonomy, in a limited number of situations. As such, it should *not* be mandatory for any kind of AI system, irrespective of the context and purpose of using it. As elaborated below, most of the systems should benefit from full automation, without HITL.

Indeed, both the ban on AI autonomy and human in the loop are *exceptions*. They should be *restricted to systems dealing with high-stakes decisions* or *high inherent risks*. Although people sometimes prefer to have a human determining outcome that affects them[98], ordinary uses of AI should benefit from full automation[99]. Ralf Poscher provides a clever summary of this argument:

"(...) refocusing on the abstract danger for concrete, substantive fundamental right's interests allows for a discussion on thresholds. Also, in the analog world, the

[96] D. LEHR – P. OHM, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, in *University of California Davis Law Review*, Davis, v. 51, n. 02, 2017, 716: «Many think that the best way to ensure fairness or justice is to inject a human into the decision-making process, perhaps with the veto power to override the inanimate counterpart. We are worried that if we simply thrust the human at the output end of the running model, there is very little she can do to root out bias. The human becomes a rubber stamp for the machine, providing nothing more than a cosmetic reason to lull ourselves into feeling better about the results. There might be better, more productive roles for human oversight elsewhere in the process».

[97] A.Z. HUQ, *Constitutional Rights in the Machine-Learning State*, in *Cornell Law Review*, Ithaca, v. 105, n. 07, 2020, 1906: «Remedies for a due process deficit are unlikely to take the form of additional human review but rather better algorithmic design».

[98] See, for instance: D.E. BAMBAUER – M. RISCH, *Worse Than Human*? in *Arizona State Law Journal*, Phoenix, v. 53, n. 04, 2021, 1091: «The surveys explore whether people prefer to have an algorithm or a human determine an outcome affecting their welfare in a range of representative scenarios with varying stakes».

[99] A conclusion endorsed by other authors: M. GUIHOT – A.F. MATTHEW – N.P. SUZOR, *Nudging robots: Innovative solutions to regulate artificial intelligence*, in *Vanderbilt Journal of Entertainment and Technology Law*, Nashville, v. 20, n. 02, 2017, 396 ss.: «The Authors propose that risk should be considered as a quality that differentiates classes of AI. (…) Each category does not and cannot justify or require the same regulatory response, and some applications may not even require a regulatory response at this stage. It is only when the risk profile of an AI application increases that a regulatory response may be required».

law does not react to each and every risk that is associated with modern society. Not every abstract risk exceeds the threshold of a fundamental rights infringement. There are general life risks that are legally moot.

(…)

The threshold for everyday life risks holds in the analog world and should hold in the digital world, too. In our digital society, we have to come to grips with a – probably dynamic – threshold of everyday digital life risks that do not constitute a fundamental rights infringement, even though personal data have been stored or processed. For AI technologies, this could mean that they can be designed and implemented such that they remain below the everyday digital life risk threshold."[100]

Having this lesson in mind, the following question is: what are the thresholds to decide if an AI system should be subject to human in the loop? It's easier to ask than to answer since there is no consensus on the matter in the scientific literature[101]. *I recommend a case-by-case analysis, considering*: 1) the average risk level of an AI system, identified by proper risk assessment (low or even medium risks should be exempted from HITL unless otherwise recommended by the risk assessment); 2) if a failure in the AI tends to compromise fundamental values or not; 3) if the system's decision is reversible or irreversible (or at least hard to reverse); 4) if the definition of "right" or "wrong" can be mathematically coded or if it is intrinsically subjective (such as in moral judgments); 5) the societal as well as the economic impacts of a failure.

Considering these factors on a case-by-case assessment one can balance the scales between technological development and innovation, on one side, and protection of fundamental rights and avoiding systemic risks, on the other side. No matter the criteria used to assess the need for a ban on AI autonomy or HITL, it is of paramount importance that regulators *clearly communicate* the adopted criteria to key stakeholders, such as developers, users, and academics.

[100] R. POSCHER, *Artificial Intelligence and the Right to Data Protection*, in *Max Planck Institute Working Paper n. 03*, 2021, 9 ss.

[101] For instance, OECD recommend the following criteria: OECD – ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. *OECD Framework for the Classification of AI Systems*. Available at: <https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en>. Access: 20 Jun. 2022 67: «Regardless of the number of risk levels or which organisation proposes them, the following are typical criteria for determining the risk level of an AI application or system Scale, *i.e.* seriousness of adverse impacts (and probability); Scope, *i.e.* breadth of application, such as the number of individuals that are or will be affected; Optionality, *i.e.* degree of choice as to whether to be subject to the effects of an AI system».

6. – The global landscape of AI regulation is still a patchwork of initiatives coming from a vast array of sources such as leading countries, companies, and international organizations[102]. This section highlights *some* important sources, *briefly* demonstrating that the ideas developed through this study are not the author's voice alone but in compliance with the international debate. Bear in mind that this section does *not* intend to dig deeper into any of these sources since each of them would demand an exclusive paper for a complete analysis. Furthermore, there are many other countries worth studying, such as China, the UK, Australia, and Uruguay. But for the sake of brevity, the author made a choice and decided to mention only the items listed below.

OECD. The Organization for Economic Co-Operation and development – OECD has published two documents of paramount importance: the AI Principles, from May 2019, and the 2022 Framework for Classifying AI Systems. Considered as a foundational document, *the OECD AI Principles*[103] was the first intergovernmental standard on AI and aim at setting *"standards for AI that are practical and flexible enough to stand the test of time"*. It comprises 5 values-based principles for the development and use of AI systems, in both the public and private sectors, irrespective of the kind of AI at stake, as well as 5 recommendations for policy makers around the world. Those principles are: *"1) inclusive growth, sustainable development and well-being; 2) human-centred values and fairness; 3) transparency and explainability; 4) robustness, security and safety; and 5) accountability"*. This study is aligned with all OECD principles and the subjects discussed herein directly relate to them, especially to principles numbers 3, 4 and 5.

Another source of paramount importance is the February 2022 *Framework for Classifying AI Systems*[104]. OECD states that it is *"a user-friendly framework for policy makers, regulators, legislators and others to characterise AI systems for specific projects and contexts. The framework links AI system characteristics with the OECD AI Principles, the first set of AI standards that governments pledged to incorporate into policy making and promote the innovative*

[102] For instance, the OECD maintains a repository with "over 700 AI policy initiatives from 60 countries": OECD – ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. *National AI policies & strategies*. Available at: < https://oecd.ai/en/dashboards>. Access: 26 Jun. 2022.

[103] For a detailed description of each principle, see: OECD – ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. *OECD AI Principles overview*. Available at: <https://oecd.ai/en/ai-principles>. Access: 20 Jun. 2022.

[104] OECD – ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. *OECD Framework for the Classification of AI Systems*. Available at: <https://www.oecd-ilibrary.org/science-and-technology/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en>. Access: 20 Jun. 2022.

and trustworthy use of AI."[105] There is a remarkable difference in scope between the 2019 Principles and the 2022 Framework. While the former is intentionally flexible to encompass all kinds of AI (as expected from a principled document), the latter was designed to focus on specific projects and contexts[106]. Those contexts (labelled as "dimensions" in the document) are: 1) people & planet; 2) economic context; 3) data & input; 4) AI model; and 5) task & output. Moreover, the framework was designed to be applicable in both "the lab" and "the field"[107]. In a near future, OECD has plans to populate the framework with the analysis of more actual systems, to develop metrics able to help assess those system's impact on human rights and well-being. A step forward would be the creation of a risk-assessment framework, in line with the ideas developed in Section 5 of this article.

UNESCO. In November 2021, the UNESCO General Conference approved a *Recommendation on the Ethics of Artificial Intelligence*[108], proposing a series of principles, many of which overlap the OECD principles, such as fairness, transparency, explainability, security, and accountability. Moreover, it recommends a continuous risk assessment for AI. Two parts of that recommendation are specially aligned with the ideas developed in this study. First, the discussion around a tradeoff between transparency and explainability, reaching out for feasible algorithms[109]. Second, despite not

[105] *Op. cit.*, 6.

[106] *Op. cit.*, 16: «The framework primary purpose is to characterise the application of an AI system deployed in a specific project and context, although some dimensions are also relevant to generic AI systems».

[107] *Op. cit.*, 7: «AI "in the lab" refers to the AI system's conception and development, before deployment. It is applicable to the Data & Input (*e.g.*, qualifying the data), AI Model (*e.g.*, training the initial model) and Task & Output dimensions (*e.g.*, for a personalisation task) of the framework. It is particularly relevant to ex ante risk-management approaches and requirements. AI "in the field" refers to the use and evolution of an AI system after deployment and is applicable to all the dimensions. It is relevant to ex post risk-management approaches and requirements».

[108] UNESCO – UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION. *UNESCO Recommendation on the Ethics of Artificial Intelligence*. Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000380455>. Access: 19 Dec. 2021, 4: «This Recommendation addresses ethical issues related to AI. It approaches AI ethics as a holistic framework of interdependent values, principles and actions that can guide societies in the AI system lifecycle, referring to human dignity and well-being as a compass to deal responsibly with the known and unknown impacts of AI systems in their interactions with human beings and their environment».

[109] *Op. cit.*, 17: «Feasibility: many AI algorithms are still not explainable; for others, explainability adds a significant implementation overhead. Until full explainability is technically possible with minimal impact on functionality, there will

quoting "human in the loop", the recommendation seems to call out States to implement a similar mechanism[110].

European Union. The EU is one of the most fructiferous sources of AI regulation. Therefore, many of their initiatives could be mentioned here. However, for the sake of brevity, only two will be addressed. Starting with the February 2020 white paper *On Artificial Intelligence: A European approach to excellence and trust*[111] which recognizes that *"Europe's current and future sustainable economic growth and societal wellbeing increasingly draws on value created by data"*[112] and that *"AI is one of the most important applications of the data economy"*[113]. This white paper aims at shaping the development and ethical use of AI inside the Union through common grounds, avoiding fragmentation among member States, and consolidating the EU as a global leader in the data-driven economy. It recommends leveraging the current industrial infrastructure and human resources and strengthening the cooperation between the Member States to build a common AI regulatory framework. The white paper has many touch points with this study. Three of them are worth mentioning. First, it lists some criteria for AI risk assessment, such as the values at stake, the inherent risk of using each system, the protection of consumers, personal data, and other fundamental rights[114]. Second, it highlights that the specific purpose of using each system is crucial to defining the expected accuracy rate[115]. Stressing that low accuracy rates should be accepted in some contexts, such as in high inherent risk situations or unhealthy activities. Third, it recommends a mandatory human in the loop for high-risk AI, leaving medium or low-risk applications

be a trade-off between the accuracy/quality of a system and its level of explainability».

[110] *Op. cit.*, 9: «It may be the case that sometimes humans would have to share control with AI systems for reasons of efficacy, but this decision to cede control in limited contexts remains that of humans, as AI systems should be researched, designed, developed, deployed, and used to assist humans in decision-making and acting, but never to replace ultimate human responsibility».

[111] EUROPEAN UNION. *On Artificial Intelligence: A European approach to excellence and trust*. Available at: <https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en>. Access: 10 May. 2020.

[112] *Op. cit.*, 1.

[113] ID.

[114] *Op. cit.*, 17.

[115] *Op. cit.*, 20: «Ensuring clear information to be provided as to the AI system's capabilities and limitations, in particular the purpose for which the systems are intended, the conditions under which they can be expected to function as intended and the expected level of accuracy in achieving the specified purpose».

outside the scope of the rule[116]. Those three aspects are totally in line with the propositions of this study.

Building on the ideas of the white paper, and after a broad consultation of the Member States, civil society, AI companies, academics, and other key stakeholders[117], the EU published a proposal of harmonized rules on artificial intelligence, called *The Artificial Intelligence Act*[118], in April 2021. The Act also adopts a risk-based approach and has similarities with many points of this research. First, it attempts to strike a balance between market innovation and the protection of fundamental rights[119]. Second, it provides mandatory requirements and *ex-ante* measures focused on high-risk systems, exempting medium and low-risk ones[120]. Third, under a title named "Prohibited Artificial Intelligence Practices", the Act imposes a ban on AI autonomy for a limited number of systems, such as credit scoring in the public sector or law enforcement based on real-time remote biometrics in public spaces, due to their "unacceptable risk"[121]. Fourth, the Act highlights that the purposes of using each system and the way it is deployed are key factors for risk assessment[122], exactly as detailed in Sections 2 and 5 of this research. Finally, the Act stresses that different problems may arise

[116] *Op. cit.*, 21: «Human oversight helps ensuring that an AI system does not undermine human autonomy or cause other adverse effects. The objective of trustworthy, ethical and human-centric AI can only be achieved by ensuring an appropriate involvement by human beings in relation to high-risk AI applications».

[117] D. HUBERT, *Initial Appraisal of a European Commission Impact Assessment*, available at: <https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2021)694212>. Access: 02 Aug. 2021, 7: «The online public consultation on the AI White Paper ran from 19 February to 14 June 2020 and received 1.215 contributions from a wide variety of stakeholders (citizens 33%, business and industry 29%, civil society 13%, academia 13% and public authorities 6%; 84% of the contributions came from Member States and the rest from outside the EU)».

[118] EUROPEAN UNION. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>. Access: 21 Jun. 2022.

[119] *Op. cit.*, 3: «(…) this proposal presents a balanced and proportionate horizontal regulatory approach to AI that is limited to the minimum necessary requirements to address the risks and problems linked to AI, without unduly constraining or hindering technological development or otherwise disproportionately increasing the cost of placing AI solutions on the market».

[120] ID.

[121] *Op. cit.*, 12.

[122] *Op. cit.*, 13: «(…) the classification as high-risk does not only depend on the function performed by the AI system, but also on the specific purpose and modalities for which that system is used».

according to the different levels of AI interference in human decision-making, such as developed in Section 4.

The United States of America. As one of the leading countries in AI research and development, the USA are a vast source of regulatory initiatives. This paper will address the Senate bill called *Algorithmic Accountability Act*[123], proposed in April 2019, and updated in February 2022. Summing up, it *"requires companies to assess the impacts of the automated systems they use and sell, creates new transparency about when and how automated systems are used, and empowers consumers to make informed choices about the automation of critical decisions."*[124] In line with the other sources mentioned above, the Algorithmic Act also focuses on high-risk systems, expressly exempting small and medium companies, since it applies to companies that have average annual gross receipts greater than US$ 50,000,000, equity value greater than US$ 250,000,000 or deals with identifying information about more than 1,000,000 people or devices[125]. One of the most controversial aspects of the bill is its possible impact on content moderation in online social networks, a topic briefly mentioned in Section 5.3 of this study.

Brazil. Although Brazil is not a leading player in the AI field, that subject is currently a hot topic in the country, with some local sources worth mentioning. The first major effort can be traced back to the *Brazilian Strategy for Digital Transformation (E-Digital)*[126] from March 2018, which aimed at harmonizing and coordinating government initiatives on digital issues in general. Even though E-Digital does not mention AI, it has laid down the foundation for future initiatives.

Following the release of the E-Digital, little had happened with AI policymaking in Brazil until two Senate bills were introduced in September and October 2019. They tried to be in line with international norms, such as the OECD Principles and intended to be complementary, targeting all kinds of AI, regardless of the economic sector, or whether the system was used by public or private entities. Coincidentally, both had only 7 articles, much shorter than average Brazilian regulations. *Bill 5,051/2019*[127] defined

[123] THE UNITED STATES OF AMERICA. *Senate – The Algorithmic Accountability Act*. Available at: <https://www.congress.gov/bill/117th-congress/house-bill/6580/text?r=2&s=1>. Access: 22 Jun. 2022.

[124] Extracted from the Senate bill summary.

[125] *Op. cit.*, 3.

[126] BRAZIL. *Decree 9,319/2018*. Available at: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/decreto/D9319.htm>. Access: 09 Feb. 2020.

[127] BRAZIL. *Senate – Bill 5,051/2019*. Available at: <https://www25.senado.leg.br/web/atividade/materias/-/materia/138790>. Access: 09 Feb. 2020.

principles for the development and use of AI while *Bill 5,691/2019*[128] proposed a national AI strategy. Due to a lack of technical and political grounds,[129] they were rapidly overcome by another legislative proposal, explained below.

The *Chamber of Deputies Bill 21/2020*[130] was presented on February 4, 2020. It is more detailed and more technical than the Senate bills, intending to overcome them. Bill 21/2020 is broadly consistent with global norms, providing for the use of AI to be based on respect for human rights and democratic values, equality, non-discrimination, plurality, transparency, autonomy, and data privacy. It also introduces a mandatory AI impact assessment. However, the bill has two major flaws. First, it does not adopt a risk-based approach, such as the international sources mentioned above, nor exempt small and medium companies from its provisions. Second, it is quite confusing at differentiating between two AI agents that the bill describes, named "development agents" and "operating agents". The former would be the entities that participate in the planning, design, data collection, processing, and construction of the AI model, as well as its verification and validation, while the latter would be entities that participate in the monitoring and operation of AI systems. Since the bill imposes different liability rules for each agent, their precise definition is a core aspect.

Meanwhile, the Executive branch moved to reclaim its leading role in AI governance, taking advantage of the delay in the legislative process of the bills to develop its own strategy, finally published in April 2021. The *Brazilian AI Strategy*[131] has nine pillars, which are grouped into three horizontal axes and six vertical axes. The three horizontal (or thematic) axes are: (i) legislation, regulation and ethical use; (ii) AI governance; and (iii) international aspects. The six vertical (or applied) axes are: (i) education; (ii) workforce and training; (iii) R&D and entrepreneurship; (iv) applications in the productive sectors; (v) applications in government; and (vi) public

[128] BRAZIL. *Senate – Bill 5,691/2019*. Available at: <https://www25.senado.leg.br/web/atividade/materias/-/materia/139586>. Access: 09 Feb. 2020.

[129] For a detailed analysis of Bill 5,051/2019, see: L. Parentoni – R.S. Valentini – T.C.O. Alves, *Panorama da Regulação da Inteligência Artificial no Brasil: com ênfase no PLS n. 5.051/2019*, in *Revista Eletrônica do Curso de Direito UFSM*, Santa Maria, v. 15, n. 2, 2020.

[130] BRAZIL. *Chamber of Deputies – Bill 21/2020*. Available at: <https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2236340>. Access: 15 Mar. 2020.

[131] BRAZIL. *Administrative Rule 4,617/2021*. Available at: <https://www.in.gov.br/en/web/dou/-/portaria-gm-n-4.617-de-6-de-abril-de-2021-*-313212172>. Access: 16 Apr. 2021.

safety. The main criticism is that the strategy is way too abstract, up to the point of being insufficient to guide practical measures.

This overview of some important global AI regulatory initiatives shows that the arguments developed through this study are not the author's voice alone but find eco in numerous international sources.

7. – There seems to be a misalignment between what AI *can* currently deliver to mankind and the results some people *expect* from it. Therein comes the central question of this study: *what should we reasonably expect from AI*? The author tried to provide a scientific and solid grounded answer to that question, contributing to a realignment of expectations around AI systems, considering their current stage, both in the lab and in the field.

To reach that answer this study developed a systematic analysis of some core factors. First, is the fact that AI is not a single, monolithic concept. On the contrary, it embraces a wide variety of applications, in different market sectors, based on a vast array of techniques and models, used for way too different purposes. Therefore, the assessment should be made *on a case-by-case basis*, considering *the actual system* and *the developers' and retailers' strategies* to introduce it in the market. After all, different purposes and strategies may implicate different kinds of risks. Second, the assessment must *identify the level of AI interference in human decision-making* since each level poses different kinds of problems and risks. To help with that, the author proposed a 3-level categorization explained through the text, as a reasoning tool for guiding the assessment of most situations.

This study then elaborates on 3 core criteria for evaluating an AI system: 1) the *accuracy rate*; 2) *level of transparency/explainability*; and 3) special situations of regulatory interference, to forbid the use of some systems due to their unacceptable risks (*ban on AI autonomy*) or at least to impose a meaningful human intervention, able to override the system's decision, if necessary be (*human in the loop*). Making it clear that *these criteria must be jointly evaluated*, since they interfere with each other. It also highlights that the excessive focus on just one of them (as the usual focus on accuracy *or* transparency) can not only compromise innovation but also curtail competitiveness and wellbeing, as the cases mentioned in the text illustrate, a point also recognized by the OECD and other international sources.

Finally, acknowledging that AI is constantly changing and evolving, this research tried to provide more abstract and time-proof criteria to set what we should reasonably expect from AI in each context. There is certainly room for further developments in the area and this study will have served its purpose by contributing to the debate.

Abstract

WHAT SHOULD WE REASONABLY EXPECT FROM ARTIFICIAL INTELLIGENCE?

L'intelligenza artificiale (o semplicemente AI) è una delle tecnologie più pervasive e all'avanguardia del nostro tempo. È già presente in diversi settori, come l'agricoltura, l'industria, il commercio, l'istruzione, i servizi professionali, le città intelligenti, la difesa informatica e così via. Tuttavia, sembra esserci un disallineamento tra ciò che i sistemi di IA possono attualmente fornire all'umanità e i risultati che alcuni si aspettano da essa. Questo disallineamento originario porta a due risultati indesiderati. In primo luogo, alcune persone si aspettano dall'IA risultati che essa, almeno nel suo attuale stadio di sviluppo, non è in grado di fornire. In secondo luogo, le persone sono insoddisfatte di ciò che l'IA è già in grado di fornire, anche se in molti contesti tali prestazioni possono essere sufficienti.

In questo articolo, l'autore sottolinea come questo disallineamento originario derivi dalla falsa premessa che i sistemi di IA debbano sempre fornire tassi di accuratezza elevati, molte volte superiori agli standard umani, indipendentemente dal contesto. Illustrando le diverse applicazioni di mercato comprese nel termine generale "IA", l'autore dimostra che non si tratta di un concetto unico e monolitico. Al contrario, l'IA abbraccia un'ampia varietà di applicazioni settoriali, ognuna delle quali ha scopi diversi, rischi intrinseci e accuratezza desiderata. L'articolo dimostra poi che ogni scopo dovrebbe puntare a un diverso tasso di accuratezza e trasparenza, caso per caso. A seconda del contesto, i sistemi di IA sono più che benvenuti, anche se alla fine forniscono tassi di accuratezza o trasparenza inferiori agli standard umani. Di conseguenza, l'autore sostiene un riallineamento delle aspettative finalizzato a 1) collegare il dibattito a ciò che l'IA è sia in laboratorio che sul campo; 2) comprendere meglio il suo potenziale e i livelli di accuratezza e trasparenza accettati in ogni contesto, considerando i diversi scopi e i rischi intrinseci dell'attività da automatizzare; e 3) fornire una guida più solida a regolatori, sviluppatori e clienti.

***

Artificial Intelligence (or just AI) is one of the most pervasive and cutting-edge technologies of our time. It is already present in a variety of sectors, such as agriculture, industry, commerce, education, professional services, smart cities, cyber defense, and so forth. However, there seems to be a misalignment between what AI systems can currently deliver to mankind and the results some people expect from it. This original misalignment leads to two unwanted outcomes. Firstly, some people expect results from AI that it – at least in its current stage of development – simply cannot deliver. Secondly, people are dissatisfied with what AI is already capable of providing, even though such provisions may be enough in many contexts.

In this article, the author points out that this original misalignment stems from the false premise that AI systems should always provide high accuracy rates, many times higher than human standards, no matter the context. By unfolding different market applications included in the general term "AI", the author demonstrates that it is not a single, monolithic concept.

On the contrary, AI embraces a wide variety of sector-specific applications, each of them with different purposes, inherent risks, and desired accuracy. The article then demonstrates that each purpose should target a different accuracy and transparency rate, on a case-by-case basis. Depending on the context, AI systems are more than welcome, even if they eventually provide accuracy or transparency rates lower than human standard. Consequently, the author advocates for a realignment of expectations fine tunned to 1) connecting the debate to what AI is both in the lab and in the field[132]; 2) better understanding its potential as well as its accepted levels of accuracy and transparency in each context, considering different purposes and inherent risks of the activity to be automated; and 3) providing more solid guidance to regulators, developers, and customers.

-----

[132] To use the same expression as the OECD, as will be seen in Section 6.