DEMOCRACY, ARTIFICIAL INTELLIGENCE AND REGULATORY CHALLENGES: RIGHTS, DILEMMAS AND POWER IN DATAFIED SOCIETIES

Sivaldo Pereira da Silva[1]

Abstract: This article aims to identify key issues in the debate on Artificial Intelligence (AI) policies today, having as premise the appropriate regulatory horizon for this growing field. The analysis was guided by documentary and bibliographic research, taking normative principles from Political Theory and Democracy Theories. After summarizing the most prominent aspects of the modus operandi of AI systems, the work listed seven key political problems that are at the foundation of this discussion: (1) personification and imputability of the machine; (2) dilemmas and moral judgments; (3) statistical authoritarianism of the metrics; (4) mathematical obscurantism in the processes; (5) utilitarian pervasiveness of autonomous systems; (6) frontiers of efficiency and control; (7) diversity and representativeness in the codes.
Keywords: Artificial Intelligence; Algorithmic governance; Philosophy of Technology; Digital Communication and Regulation; Public policies

1 Introduction

The innovations provided by systems and machines based on Artificial Intelligence (AI) are today a relevant boost in the efficiency of processes in the most diverse areas such as communication, politics, transport, public safety, health, education, etc. The tendency, for the next decades, is that there will be significant leaps in this field, with the horizon of becoming increasingly ordinary and omnipresent. These technologies mean much more than purely instrumental tools. They are, in fact, technical-cultural artifacts that substantially alter decision-making processes, with effects on different branches of human activity, from consumption parameters to the power relations between different actors (whether between State and citizens; companies and consumers, organizations and individuals).

To deal with these changes, governments and organizations in various jurisdictions at the local, national and multilateral levels (such as New York City, the German Federation, the European Union, the UN, OECD) are developing strategic plans, legislation or public policies that aim to resolve the tensions arising from this situation, as well as to welcome such innovations to guarantee their potential positive effects.

[1] Professor at the College of Communication (FAC) and the Graduate Program in Communication at the University of Brasília (UnB). PhD in Contemporary Communication and Culture by the Federal University of Bahia, with a doctoral internship at the University of Washington (USA). He was a visiting researcher at the Institute of Applied Economic Research (IPEA); *ad hoc* consultant at Unesco for the application of media development indicators in Brazil. He is the founder and coordinator of the research group Research Center for Communication, Technology and Politics (CTPol) and researcher at the National Institute of Science and Technology in Digital Democracy (INCT-DD).

In addition to the practical benefits that the most advanced algorithmic systems provide, this also tends to generate new forms of inequality, violations of rights or increase the concentration of power. The proliferation of autonomous systems has repercussions on politically sensitive points such as privacy, freedom, individual and collective rights, misinformation, ethical transgressions, authoritarianism etc.

Given this context, this work aims to identify and characterize the main key problems that any Artificial Intelligence policy needs to answer. In this sense, the article brings an exploratory study, based on documentary and bibliographic research, analyzed under the normative lens of democratic principles. For this, the study is divided into two parts: first, it makes a conceptual approach on Artificial Intelligence (AI), its origins and fundamental characteristics. The second part summarizes seven important key problems of political emphasis that are at the basis of the current regulatory debate on AI in the world and that are decisive discussions for understanding the complex relationship between democracy and digital autonomous systems.

2 Artificial intelligence: technique beyond the technique

The expression "Artificial Intelligence" leads us to imagine thinking machines or self-conscious technical artifacts. However, as occurs in every metaphor, it is a generalist terminology that helps us cognitively in a synthesis of the phenomenon, but minimizes important aspects, generating a definition that lacks more precision. In any case, we should not consider the use of this terminology as a problem, because it is already widespread in the social imaginary, in government documents, news, company guidelines, etc. It is possible to adopt it as long as we can contextualize and dimension this metaphor and highlight, above all, the elements that the expression hides.

Specifically, Artificial Intelligence refers to a set of logical methods that aim to solve problems based on trained algorithms (through **inputs**, data entry) to understand patterns, learn from errors and reconfigure themselves reaching results (**output**) each time closer to what is expected. Therefore, it is important to note that we are not talking about a machine that thinks, but that solves logical problems and is trained in this sense from the experience (data) it receives.

From the historical point of view, technical artifacts that helped in the realization of some logical operation, especially mathematics, are not new. In several cultures, such as Mesopotamia and China, instruments such as the abacus have existed since antiquity. In modern times, these mechanisms gained a new version with the first calculators. As Gleick puts it (2011, p.99):

> Blaise Pascal created an adding machine in 1642, with a row of rotating disks, one for each decimal digit. Three decades later, Leibniz improved Pascal's work by using a drum with protruding teeth to "regroup" the units from one digit to the next.

However, the author recalls that the prototypes of Pascal and Leibniz remained very close to the abacus, as they made passive records of the memory` states of a particular mathematical operation.

In the 19th century, in the context of the Industrial Revolution, Charles Babbage took a step further by inserting an important element in calculating machines: automatism. This set a precedent in the development of the computer that would be actually created in the following century. However, Babbage's machine was mechanical (it did not use electricity) and did not presuppose a versatile logical structure, such as a binary perspective (based on two digits, 0 and 1) or Boolean variables (true/false). The ground for Artificial Intelligence as we know it today actually starts to become more fertile in the second half of the 20th century. More specifically, its origins are linked to the term "**machine intelligence**" disseminated by Alan Turing, with first records in manuscripts still in 1941 (COPELAND, 2004) [2].

The basic principle of Turing's idea concerned the solution of logical and mathematical problems through automation in binary electronic systems and the possibility of building machines capable of learning from experience. In an article published in 1950 entitled "*Computing machinery and intelligence"* in *Psychology and Philosophy* magazine, Turing proposes to think about the following question: "Can machines think?" (Turing, 1950). Although the author has a long discussion of the objections to this question, his concern is actually to reformulate such questioning in the direction of what he called the "Imitation Game". For the author, the universal imitation capacity of a machine (based on binary language) would be one of the differential elements that deserved special attention:

> This special property of digital computers, that they can mimic any discrete state machine, is described by saying that they are universal machines. The existence of machines with this property has the important consequence that, considerations of speed apart, it is unnecessary to design various new machines to do various computing processes (TURING, 1950, p. 441).

In 1943 Warren McCulloch and Walter Pitts published an article entitled "A logical calculus *of the ideas immanent in nervous activity*" explaining how neurons work and modeled a simple artificial neural network using electrical circuits to demonstrate their hypotheses (MCCULLOCH; PITTS, 1943)[3]. Adding this to Turing's ideas, several researchers were

[2] Although Vannevar Bush did not speak directly about artificial intelligence, in 1945 he published an article entitled "As we may think" in The Atlantic Monthly magazine. He proposed a collective memory machine that he called Memex, capable of agglutinating and processing information, transforming it into knowledge. A reproduction of this text is available at: <https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/> Access on July 4, 2019.

[3] Other works also contributed in this line: in 1949 the book "*The organization of behavior: a neuropsychological*

encouraged to think about artificial neural networks in computers and how this could be aligned with the design of an "intelligent machine". In 1956, in the state of New Hampshire (USA) the term “Artificial Intelligence” as we know it today appeared for the first time in a conference entitled “*The Dartmouth Summer Research Project on Artificial Intelligence*”, considered by many as the founding stone of this field of search.

However, if the notion of Artificial Intelligence existed at least half a century ago, why are we only now talking about laws and regulations for this field as if it were something discovered recently? The explanation is relatively simple: because a set of technical conditions that were not given before, started to co-exist and converge mainly from the first decades of this century[4]. In this scenario, it is estimated that there will be a boom in the use of AI in the next two or three decades (CATH, 2017; DAFOE, 2018; GRACE et al, 2018).

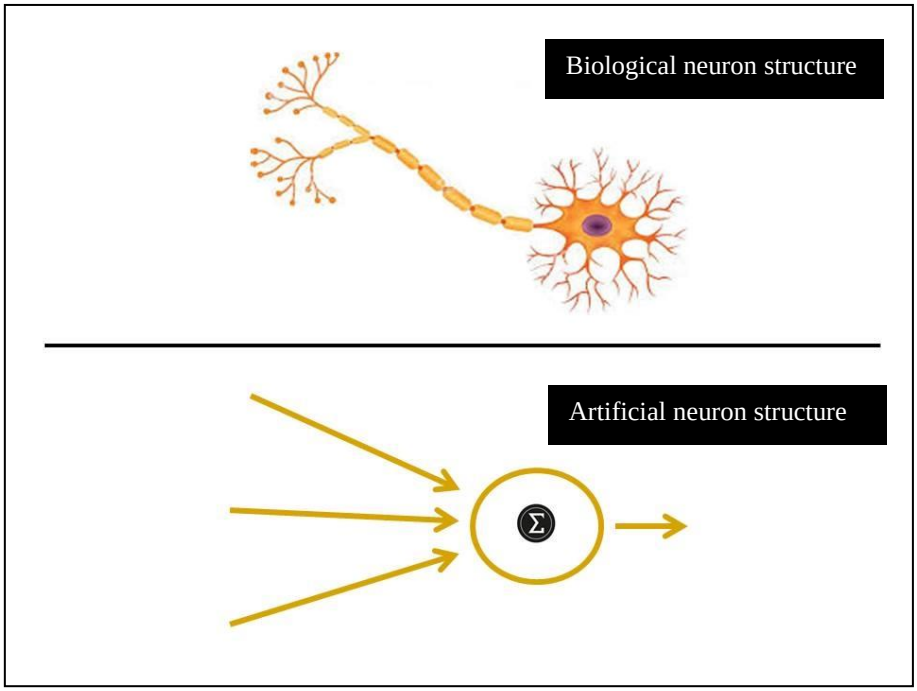
To better understand the nature of this phenomenon, it is convenient to summarize four aspects that deserve special attention because they represent dimensions that help us to understand the functioning of systems based on Artificial Intelligence: (a) artificial neural networks; (b) the meaning of the notion of imitation; (c) the power of automation and (d) the typological levels of AI.

First, for a system to be called "intelligent", it presupposes that it is capable of learning and making decisions based on logic. An innovative method in this sense is the idea of artificial neural networks (also a metaphor linked to the brain) that has become one of the most promising and influential conceptions of AI, on which the techniques of machine learning and deep learning are based. More didactically, an artificial neural network is a composition of algorithms inspired by the structure and functioning of a biological neuron, as shown in Figure 1.

theory", by Donald Hebb; and in 1958 Frank Rosenblat’s article entitled "*The perceptron: a probabilistic model for information storage and organization in the brain*".

[4] With the creation of advanced digital infrastructures, especially 5G; intensification of Big Data mechanisms that reflect the exponential capacity to collect and process large volumes of data, from various sources with speed that did not exist before; development of more sophisticated algorithms, etc.

Figure 1 - Comparative illustration of the biological and artificial neuron structures**.**



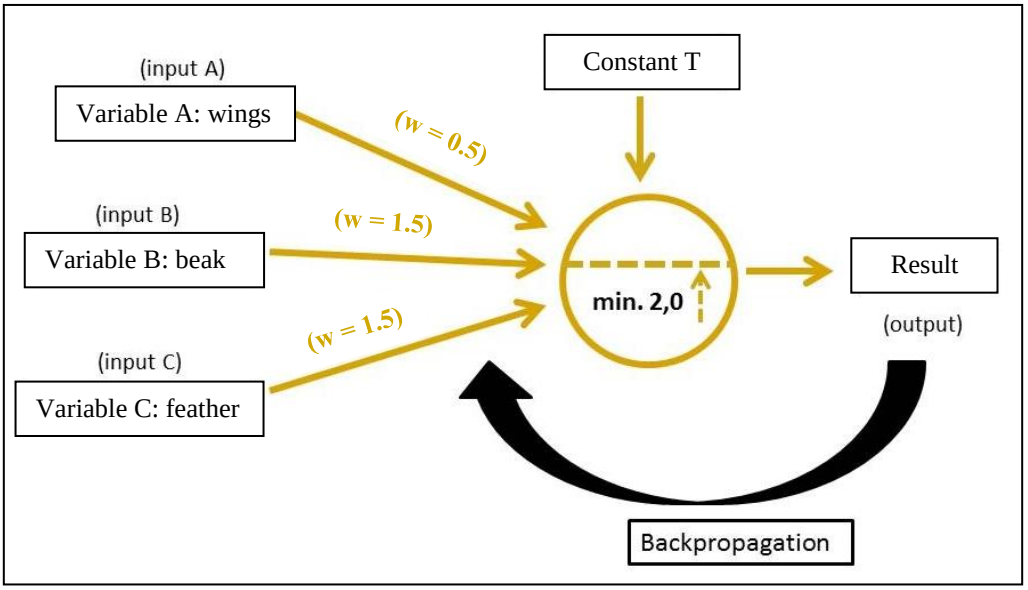
Source: self-elaboration based on the proposal by McCulloch and Water Pitts (1943)

Basically, an artificial neuron is made up of several inputs designed to capture information from the several variables that define what is being evaluated (for example, taking the variables "beak" or "wings" as relevant indicators to define whether an image is or not referring to a bird). For each input (variables captured) a weight is given. The quantitative of the values of each variable with their respective weights is added to a single value which is then subjected to a nucleus in which it is required that a minimum quantity is reached for the neuron to be activated (such as a trigger that will only be activated if the sum reaches a value considered significant)[5]. Thus, for the neuron to respond positively that the image in question is of a bird, there must be a combination of the variables and their respective weights with values that reach a certain level capable of indicating that there is a great probability that the image is, in fact, of a bird. Figure 2 shows a simplified illustration of this process and its weights[6], taking as an example only 1 artificial neuron that is based on 3 variables (wings, feather and beak) to determine if a particular image is of a bird.

[5] This is the basic structure of an artificial neuron. With the advancement of studies in this field, other more complex elements were added to improve the accuracy of the responses.

[6] The weight values in Figure 1 are for illustrative purposes only. In practice, weights are stipulated in the programming process considering the probability of the variables.

Figure 2 - Simplified illustration of the main functioning mechanisms of an artificial neuron



Source: self-elaboration, synthesizing and adapting, for didactic purposes, the initial structure proposed by McCulloch and Water Pitts (1943)

If the neuron detects only variable A (wings) as positive and the others are negative (which would result in zero, for not being confirmed) the sum in the nucleus will be only 0.5 points (weight value of the variable "wing"), thus not reaching the minimum level stipulated in this neuron (2 points) to consider the image as that of a bird. This is because having wings is not enough to determine that something is, in fact, a bird because there are other objects that have wings and that are not birds, such as an airplane, a bat or a butterfly. In another situation, if the neuron detects variable A (wings) and B (beak) as positive, this would add up to 2 points (0.5 from variable A + 1.5 from variable B), thus reaching the required minimum score. In view of this, the neuron would be activated to give the positive response that the image is of a bird.

This is because if something has wings and beak, the possibility of being a bird is statistically increased. However, the process does not end here. What the algorithm has accomplished so far is a "bet". What actually makes the machine “learn” are the corrections of the weights after checking whether the bet has proved to be “false” or “true”. The response of the neuron is compared to reality to see if there was a right or wrong estimate.

This information (of right or wrong) generates a wave of modifications or reinforcements of the weights called "backpropagation": if a neuron (with certain weight configurations) made an error in its image estimation because it detected that it was actually the image of a plane, this error when perceived will turn into a backpropagation that will decrease the weights of these variables because they proved to be ineffective. The same is done for variables that were initially judged unimportant, but at the results proved to be decisive: this time, back propagation acts to positively calibrate these variables, emphasizing them as

relevant. The difference in the final result value is reprocessed, in this case, increasing the weights of the underestimated variables, making the neuron work correctly with the weights duly calibrated to get it right, based on previous experiences.

The sophistication of machine learning is nothing more than a correction of weights initially estimated from the found errors, thus increasing the possibility of getting it right the next time it encounter the same variables. However, for the system to work effectively, there must be many interconnected artificial neurons and, mainly, a lot of data so that the neurons can "test" the weights of the variables, that is, so that the neural networks are "trained". The so-called “training phase” of AI algorithms requires a large amount of information and, as we have seen, the availability of previous data and statistical occurrence is a determining element. For example, AI software will only be able to distinguish between “birds” and “airplanes” when it has received many input images until it can make the distinction in an automated way. Of course, the given example is only didactic. Real neural networks include hundreds or thousands of neurons with hundreds or thousands of variables being tested and recalibrated.

A second element that is fundamental to the functioning of Artificial Intelligence systems, as mentioned, is the perspective of imitation. With machine learning techniques, the algorithms have an immense capacity to identify (and repeat) patterns, enforcing the Turing game (Turing, 1950). It does not mean that the algorithm is aware of the difference between a bird and an airplane, but it does imitate our perception of things by giving weights to certain variables that are tested and defined as determinants, just as we do when looking at the elements that make up our definition of "Bird". In this sense, imitation is based on statistical probabilities where what is majority (after the back propagation process) is perceived by the algorithms and reinforced by them. In contrast, that which deviates from the prevailing pattern tends to be overlooked and ignored or becomes an outdated information in the system.

For example, if a bird appears without a beak and without wings, the algorithm will find it difficult to consider it a bird, as it deviates from the pattern that the training phase established in the code about what a bird is. Therefore, birds with congenital or injured disabilities tend not to be recognized as birds. Note that the training phase is dynamic, but after this stage, systems tend to become stable (relatively rigid) based on a majority perspective. Here we can see that the imitation property is based on something that is already given, that is, to imitate is to preserve and reinforce something pre-existing. In this sense, it is possible to affirm that there is paradoxically a mixture of innovation and conservatism in the sophistication of learning algorithms.

The third characteristic that we must keep in mind when designing Artificial Intelligence is automation. Two centuries later, Babbage's dream of searching for the automatic machine was carried out to the extreme. Automation is a central element in any AI system because machines are only considered intelligent if they are able to operate on their own from

an initial start and find their own way. The evolution in energy production and storage combined with the ability of algorithms to generate "*looping*" or *ad infinitum* repetition cycles is an important combination that ultimately implies a new form of power. A computer or system can run repeatedly for years and centuries, as long as it has power. In a world with an increasingly dated daily life, we have, in practice, the increased power of certain actors (such as the State, institutions and corporations) due to the ability to impose the repetition of procedures or forms of behavior. Authoritarianisms or unjust actions can be repeated in a much more agile way, at low cost and in a much more difficult way to be opposed when executed by autonomous systems. This can also bring greater rigidity in the relationship between asymmetric parts where the system tends to follow procedures and not observe exceptional situations. It also places the machine as a decision-making entity, and behind an apparently technical decision is the value embedded in the metrics.

Finally, a fourth important aspect to understand Artificial Intelligence is to realize that there are different degrees of development of these technologies and this has repercussions in different dimensions on the place of the artifacts. The literature has pointed to three levels or types of AI, as summarized by Girasa (2020): **Artificial Narrow Intelligence, Artificial General Intelligence and Artificial Super intelligence**. The first refers to the performance of a singular task. The second manages to perform several tasks at the same time in a similar way to the human brain. The third consists of overcoming human capacity in several aspects. We are currently in the first stage, but already with promising horizons and prototypes of second level systems. In relation to the third level, this will only be possible with the creation of more robust processing structures than the one we currently have, such as quantum computing, which is still at a very early stage of development. What is important to note in these levels is precisely the sophistication, breaking capacity and scope of action that they represent. The greater the degree of AI development, the more intense its pervasiveness and power, tending to be far more culturally and socially shocking and much more politically disruptive.

All these issues that characterize the functioning of Artificial Intelligence must be thought from parameters that manage to go beyond the horizon of technical efficiency. There are social, cultural, political and economic issues involved. A good metaphor for this is the image that the German philosopher Martin Heidegger described when he analyzed the essence of modern technology based on the Gestell concept (HEIDEGGER, 2001). For the author, in the past, we built bridges that were technical devices installed on rivers. With the use of techno science in its incessant quest to extract, manipulate and store energy, the hydroelectric power plant is not an element that is installed in the river, as the bridge was. For him, the situation was reversed: now the river is installed in the plant (because it submits the river to its objectives) and thus, the river has become a device of the technological system. Bringing this to the advances in AI systems, we are talking directly about problems of agency and human autonomy.

That said, given the expansion of these systems and their growing centrality in daily life, it is necessary to develop strategies so that the State is able to promote and stimulate all the benefits of an AI and, at the same time, mitigate its possible distortions, defining roles for that the various actors involved can interact harmoniously, guaranteeing the protection of rights; avoiding loss of autonomy and freedoms.

3 Artificial intelligence, regulation and democracy: seven key problems

Between 2017 and 2019 several countries, on all continents, launched their strategies for Artificial Intelligence: Germany, Canada, China, Denmark, United Arab Emirates, United States of America, Finland, France, India, Italy, Japan, Malaysia, Mexico, New Zealand, Kenya, Singapore, South Korea, Sweden, Taiwan, United Kingdom. Other countries that have not yet launched an official strategy (such as Australia, Spain, Poland and Uruguay[7]) were creating committees, public consultations or projecting a specific budget for the development of this area.

Regional blocks or pan-regional articulations have also been concerned with the topic, generating documents such as the *Declaration on AI in the Nordic-Baltic Region*[8] (published by a joint of Nordic and Baltic countries) or creating instances such as *High-Level Expert Group on Artificial Intelligence* (AI HLEG)[9] of the European Union. In addition, traditional multilateral organizations such as the UN[10] and OECD[11] have actions, guidelines or recommendations on the topic. Professional organizations such as the Institute of Electrical and Electronics Engineers (IEEE) and civilian articulations such as the Montreal Declaration[12] are also initiatives concerned with establishing principles for good AI practice.

This scenario demonstrates that there is an evident effervescence in the elaboration of guidelines and strategies for AI all over the world, but more concrete regulatory actions, in the form of laws, are still in process or are incipient, localized or partial (as is the case from New York City that created one of the first laws in the world on the subject and Germany that approved a law that addresses part of the problem, specifically regulating the use of autonomous vehicles).

---

[7] In the Brazilian case, there is still no defined strategy. The country has a Digital Strategy (E-digital) launched in 2018 that contains generic guidelines for digital transformation, but had not presented, until the first semester of 2020, a more specific document or official guideline for AI.

[8] https://www.regeringen.se/49a602/globalassets/regeringen/dokument/naringsdepartementet/20180514_nmr_deklaration-slutlig-webb.pdf Access on July 6, 2019.

[9] https://ec.europa.eu/digital-single-market/en/artificial-intelligence

[10] The UN has some AI initiatives, one of which is the platform "AI for Good https://aiforgood.itu.int/based on annual meetings on the topic (AI for Good Global Summit), maintained by the International Telecommunication Union (ITU). There are also other initiatives such as the Center for Artificial Intelligence and Robotics http://www.unicri.it/in_focus/on/UNICRI_Centre_Artificial_Robotics (linked to UNICRI), in addition to documents on the topic: https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf Access on January 8, 2020

[11] https://www.oecd.org/going-digital/ai/Access on July 11, 2019.

[12] https://www.montrealdeclaration-responsibleai.com/the-declaration Access on July 15, 2019.

Although incipient AI strategies and policies have their peculiarities and emphases (some with a greater focus on economic aspects, others on ethical issues), when analyzing the set of strategies or proto-regulations, it is possible to identify seven key political problems that have direct implications for contemporary democracies that deserve special attention, due to their possible consequences and political impacts. We can summarize them in the following terms: (1) personification and imputability of the machine; (2) dilemmas and moral judgments; (3) statistical authoritarianism of the metrics; (4) mathematical obscurantism in the processes; (5) utilitarian pervasiveness of autonomous systems; (6) frontiers of efficiency and control; (7) diversity and representativeness in the codes.

4 About personification and imputability of the machine

In science fiction, mainly of the dystopian type, the problem of the personification of Artificial Intelligence systems is a recurring and often central element in the plot it develops. In the 2000s, the television series *Battlestar Galactica* dramatized the difficulty of distinguishing human beings from artificial beings. In the real world, this indistinction is still not as evident and as advanced as in fiction, but this is already a problem that is beginning to emerge in the regulatory horizon of AI systems. Especially since the more AI algorithms are trained (with data from the coming years and decades), the more credible to a human interlocutor it will be, interacting with the public and tending to become increasingly difficult to be perceived as a machine. Therefore, the first question that arises with the personification of these systems is the individual's right to know whether he is actually talking to another human being or whether it consists of an AI system. This is because the communication contract that is established is quite different when we speak with machines and not with human subjectivities directly. This differentiation becomes important in the regulatory debate, as it requires rules that compel the machine to declare itself as a machine. For example, when a company or government agency adopts an AI system to serve the public, the conversation needs to start with identifying the type of interlocutor on the other end of the line or that some information in this sense is clear and evident.

The question of the personification of the machine also tends to cross the very dynamics of the political system. A problem now emerging in election campaigns is *chatbots*: algorithms that simulate conversation in natural language, manufactured and ordered (on an industrial scale) to act massively in negative campaigns or in discursive engagements in favor of a candidate. With the evolution of AI systems, the phenomenon of *chatbots* tends to be more and more common and sophisticated, spreading to several areas whether in the form of customer service systems, legal advice or even conversation services for emotional support . The question is, how to regulate this new “entity” that presents itself among us and the limits for its use

without this implying in transgressions of rights?

This also raises hypotheses about the legal personality of AI systems, for example, for some analysts, due to the complexity of these systems, it might be appropriate for a robot to have a personality similar to a company that is classified as a "legal entity" as opposed to "individual entity", that is, to the natural individual. This perspective appears in different ways in different government strategies. A good example is a resolution of the European Parliament of February 2017 entitled "*Civil Law Rules on Robotics*" that brought a series of recommendations to the European Commission to think about the place of robots in a society increasingly permeated by freelancers based on AI. In paragraph 59, letter "f", parliamentarians recommend:

> Creating a specific **legal status for robots** in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently

The resolution was not without controversy and a group of experts even published an open letter to the European Commission criticizing it suggesting ignoring such proposition[13]. However, this is still an open debate, because due to the complexity of the actors involved and the increasing degree of personification of AI systems – which break with the notion of authorship because the automaton itself is transformed by learning from its experience in the world – analysts explain that:

> [...] a legal conflict arises, as within the framework of the current legislation the robot cannot be held liable for actions and (or) inaction and as a result responsibility falls on the user, software developer, or manufacturer. At the same time, the EU resolution raises the issue of responsibility in the event that the robot caused damage due to the decisions made by the robot itself (based on the embedded algorithms) and the definition of the third party responsible for paying compensation is impossible (ATABEKOV; YASTREBOV, 2018 p. 779).

The same original code can take on different "personalities" and act differently if the algorithm is trained with data from a religious population or if it is trained with data from a population of atheists. The argument used is that the code's authors cannot be blamed if an AI system became fascist after being appropriated and trained by fascist groups. However, it is necessary to create regulatory mechanisms that oblige programmers to develop technical solutions capable of inhibiting an AI system from becoming harmful in the hands of others. It is necessary to establish degrees of responsibility for each agent in the complex production and consumption chain that goes through this type of tool.

[13] The open letter can be accessed at https://g8fip1kplyr33r3krz5b97d1-wpengine.netdna-ssl.com/wp-content/uploads/2018/04/RoboticsOpenLetter.pdf Access on August 22, 2019.

5 About dilemmas and moral judgments

One of the most important themes in theories of democracy is moral dilemmas. On the other hand, the moral dimension has also always been present in the most philosophical or fictional conception of Artificial Intelligence, whose most well-known assertion are is the principles idealized by the writer Isaac Asimov (known as the three laws of robotics). Moral judgments are relevant indicators for analyzing how democracies treat divergences, respecting differences and maintaining a rational and fair procedure in decision-making processes; and what is the role of institutions (or the State) to deliberately deal with complex conflicts that involve fundamental moral disagreements (GUTMANN; THOMPSON, 1996). Moral judgments reflect social tension and the possible imbalance in the balance of power between groups of divergent worldviews. At the same time, technical artifacts also place us before moral problems.

The trolley problem, a critical ethical experiment idealized by Philippa Foot, is a good example. Briefly, the question brings a hypothetical situation in which there is a train on a collision course that will strike five people who are on the tracks. However, an observer in a privileged position realizes the future accident and has in his hands the possibility of pulling a lever and diverting the vehicle to another alternative path. This will save the lives of five people, but the diversion will result in the death of another person who is on the diverted path. Given this hypothetical situation, the question is: would it be ethical to sacrifice one person's life to save the lives of five? Usually, most people believe that the answer is yes. On the other hand, the dilemma becomes more complex with a small variation of history: when the same train continues on the five-person collision course, however in this second version the observer is on a bridge under which the train will pass and on its side is a man who, if thrown down the viaduct in the train line, his body would serve as a brake and stop the train before reaching the people on the track, thus saving the five lives previously threatened.

The same question is asked for this second version: would it be ethical to sacrifice one person's life to save the lives of five? In this new narrative, the result is the same: sacrificing one person to save five. However, people find it more difficult to take a position in this case because there is no intermediate technical element as in the first situation: a lever that makes the act of sacrificing one life for five “less personal”.

In the current context of expansion of ubiquitous and predictive digital systems spread throughout everyday life, the Artificial Intelligence artifact occupies the privileged observer position of the “trolley problem” that foresees the future and has technical mechanisms to make decisions about lives. Faced with the advancement of the Internet of Things and the daily use of various types of autonomous machines (such as self-driving cars; autopilot drones; robots, etc.) these objects tend to be configured as new “actors” in the fauna of social dynamics and , at the

same time, will mean more and more important decision-making bodies. For example, a self-driving AI-based car makes decisions about its route on public roads at all times (if it turns left, stop at the red traffic light or continue straight on an avenue). Some decisions are obvious and expected: do not go ahead at the red light, for example. However, other decisions will have a greater degree of complexity and would be difficult even for a human being because they are moral decisions. Suppose there is a self-driving car with a couple of young passengers being transported on a road. Unexpectedly, the truck in front of this vehicle stops abruptly enough that the brake of the vehicle behind cannot respond to the point of avoiding a collision.

The accident would be fatal to the young passenger couple and there would be total loss of the vehicle. The AI system of the self-driving car will be able to anticipate the problem and will be able to make a decision on what it will do to result in the least possible damage. The problem starts when we add other elements in this equation (similar to the "trolley problem") that make the decision making process a moral position that expresses a certain worldview or values. Suppose that if the car chooses to veer off to the right, it will hit a child coming out of school who will surely die on impact, but damage to the vehicle and its passengers will be minimal and they will survive. If it chooses to turn left, it will hit a car with three elderly people who are on the opposite road and, due to the angle of the collision, the elderly will suffer the greatest impact and will certainly die while the car with the young couple will have smaller damage and they will survive.

Here we have a moral decision that needs to be made by the algorithm. If the company that built the vehicle designed the code so that the option in these cases is the least possible damage to the property (the car) and its client (the young couple), it could mean the death of a child or the death of three elderly people instead of the death of the young couple who own the vehicle. If the criterion is the smallest number of victims possible, the child on the sidewalk will be sacrificed (and thus preserve the lives of the two young people in the car or the three elderly people on the other vehicle).

If the criterion is different and considers the preservation of the lives of younger people (who would have a whole life ahead of them) as a priority in relation to the lives of older people (who have already had the opportunity to have a long life experience) the result will be death of the three elderly people on the other side of the road. Whatever the decision, it will be controversial and fall into a moral dilemma. In theory, depending on the vision of the applied regulation, the law may compel the company to choose, in this case, for the damage to the car and the passenger, preserving lives outside the accident. However, the situations can be even more problematic: suppose that, instead of a car with a young couple, the collision occurred with a school transport vehicle with 12 children.

Should the algorithm opt for the death of these 12 children who are directly involved in the accident on the road or should it sacrifice that child on the sidewalk (unrelated to the event on the road) or even the three elderly people on the opposite road? The only certainty in this case is that the parameters for this decision cannot be left only to the company that designed the codes. At least until 2016, the option of companies like Mercedes-Benz is to prioritize the life of the owner in case of moral dilemmas[14]. One of the biggest challenges is precisely how to regulate these new decision-making bodies without this representing the preponderance of specific interests and values (such as the commercial interests of the company that built the car or the interests of a group that violate the rights of third parties).

6 About statistical authoritarianism of the metrics

The automation of AI systems as well as their effectiveness are based on statistical analyses fed by a large volume of data and guided by metrics or criteria previously established by the programmer. The metrics precede the weights, that is, before the system can work, it is necessary to establish what its goals are, where it wants to go and what really matters. As Bigonha illustrates (2018, p.6):

> Consider, for example, a model that grants credit and predicts a person's score, given the likelihood that they will repay the loan. What does the success of the system represent? More profit? A higher number of people getting a loan? Higher number of payers?

In this case, whoever has the power to define the metric of the algorithm will be able to exercise new forms of encoded power in digital systems. In addition, this can be quite partial, to the point of generating discrimination, accentuating inequalities and even flirting with a new form of authoritarianism configured in the machines.

Metrics can also be authoritarian when based on information that, while having relevant statistical value and unveiling realistic prognoses, cannot be used for a given decision making because they go beyond their function or violate rights. In big data analysis, it is common to ignore the reasons for certain correlations between variables, even though it is known that they exist and such knowledge is used in decision-making processes (MAYER-SCHONBERGER; CUKIER, 2013). This causes data of the most varied types and sources to be crossed statistically so that the system makes logical-statistical, but authoritarian, decisions. For example:

[14] See in <https://www.tecmundo.com.br/mercedes/110591-sim-carro-autonomo-mercedes-atropelar-voce-salvar-condutor.htm > Access on October 20, 2016.

> [...] it is interesting to mention the controversy in Germany involving SCHUFA (a German company that provides credit protection services), which, in the context of consumer risk assessment, classified your request for access to your own data as a negative criterion. This is due to a statistical correlation that was established in the sense that consumers who accessed their score more were more likely to be in default. The company suffered numerous criticisms due to this conduct, which penalized those who wanted to contract a credit with a lower scoring, exclusively, due to the exercise of a right. (DONEDA et al, 2018, p. 6)

In this case, the regulation should impose limits on the statistical crossings of the AI systems for the use of certain information that does not concern the topic in question or that may subsidize punitive actions simply because someone exercises a right but that generated data that can turn against the individual. The challenge is how to establish these limits without creating a generic obstacle to the use of correlation statistics when crossing different information and variables.

It is necessary to define in what situations this can be considered normal and, at the same time, in what situations it becomes an abuse or violation. Requirement of documents that clearly detail how the codes were built and which metrics were used; development of deontological codes that establish guidelines and limits for companies and programmers; regular application of independent algorithmic audits, among others, are some elements that the regulatory process can use.

7 About mathematical obscurantism in the processes

One of the most recurring questions in AI policy or regulation strategies and documents is the problem of opaque algorithms. Based on the premise that algorithms are increasingly ubiquitous in the most diverse areas of human activity, having life crossed (and often determined) by rules and parameters that we do not know becomes a problem of subjects' autonomy and self-determination. For this reason, in general, the requirement for some level of control, transparency and accountability from companies that manage AI projects has been a keynote present in the documents and strategies for this field.

In a study that addressed this problem of lack of transparency, Burrell (2016) identified three types of opacities that occur around the algorithms. (a) First, **opacity as intentional corporate or state secrecy**. Companies tend to understand codes as a commercial asset and protect them as industrial secrets, keeping them out of the public sight, under the claim of copyright. Governments also classify certain codes as state secret or make their publication difficult. (b) Second, **opacity as technical illiteracy**. Algorithms are codes whose transparency and understanding is restricted to those who have knowledge of the language to decipher them. And only a tiny fraction of the population has this "literacy". (c) Third, **opacity as a structural feature of machine learning algorithms**. This last item deals with a structural problem directly linked to Artificial Intelligence, because the dense training of neural networks and their weight

calibrations makes the reconstitution of the entire process of how the system reached a certain decision difficult to visualize because the more complex the structure of neural networks, the more obscure the process becomes (LEESE, 2014; MITTELSTADT et al 2016). As Rendtorff-Smith points out (2018, p.11):

> These systems, particularly when we talk of unsupervised learning systems, are, by nature, obscure, which makes it difficult to maintain fundamental principles of transparency, explicability and accountability. Another challenge is that governments generally acquire and deploy proprietary AI technology developed by actors in the private industry.

For the author, it is important that AI systems are auditable, allowing researchers to identify the source of an error or adverse consequence. However, traditional audits in the format we know today can be ineffective because, to the naked eye, auditors will find it very difficult to decipher the dense process generated by artificial neural networks.

From a regulatory point of view, part of the solution to this technical obscurantism is the investment in research and AI tools capable of producing automated audits appropriate to deal with the deep layers of artificial neural networks. For this, it is necessary to develop technologies aimed specifically for this purpose, which can be made possible by specific organizations or agencies to stimulate and promote the development of this field. The creation of specific regulatory agencies has been an element that appears in this debate as a necessary mechanism for this and other problems of enforcement and inspection with the AI systems (SCHERER, 2016; TUTT, 2017).

8 About the utilitarian pervasiveness of autonomous systems

A fourth problem that runs through regulatory strategies and policies for Artificial Intelligence, and which deserves attention, is the pervasive and utilitarian character inherent in these technologies. Pervasive because, not so far from now, AI systems will be inserted in all areas of human experience, acting on the most diverse types of services; intermediating even our use of other devices and machines (such as automobiles, airplanes and houses); crossing social, political, cultural, economic and even affective relationships; arriving first where human beings have not been yet (as in space missions and planet colonization). Utilitarian because there is a Bentham's assumption in the widespread use and dissemination of these technologies that are capable of involving us ontologically, as there is a deep relationship between technical efficiency, comfort and well-being. Let us look at what Bentham said in the 18th century and think about how it relates to Artificial Intelligence devices today:

> By utility is meant that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness, (all this in the present case comes to the same thing) or (what comes again to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered: if that party be the community in general, then the happiness of the community: if a particular individual, then the happiness of that individual (BENTHAM, p. 14-15, [1781], 2000)

It clear that utilitarian morality is very much in line with trends in the functioning of Artificial Intelligence systems. Above all, because the exaltation discourse in strategies, products and services in this field is about the possibility of improving life in the near future for as many people as possible and the implicit majority in the metrics of artificial neural networks, as we saw earlier, reinforce this perspective. For example, when an AI system captures the majority desire of a given population, generating specific benefits suitable for this majority, those that were outside the statistical standards tend to be considered deviant, so the exclusion generated by the artifact would be justified on the basis of a utilitarian morality that was able to generate happiness for as many people as possible.

At the same time, as Helbing and colleagues (2017) point out, these technologies will be able to connect more and more, enter our lives and follow the entire path of our actions and the tendency is that "computer programming moves towards people programming" who will gradually be linked and dependent on these systems, partly by choice (by considering it more comfortable and efficient), partly because there is cultural and social pressure, typical of datafied societies:

> It is possible to realize that in an increasingly dated world with an abundance of offers and massive use of logical applications to help in routines, not to give up data and decide not to feed dataveillance (or digital surveillance) means opting for a life of exclusion from the diverse amenities that the system offers. It will imply a harder life, with less comfort, whose execution of routines will require more time and energy to be performed. In the end, the great problem of privacy today is linked to the practical-logical bias of digital culture, which ultimately means better living, that is, with more comfort and practicality, even if with less autonomy (SILVA, 2019, p. 164).

To make matters worse, Nemitz (2018) points out that the pervasiveness of AI systems will govern the bulk of a society's functions (health, education, security, economy, justice etc.) tending to remain in the hands of a few companies that will in fact concentrate the necessary know-how and infrastructure. And this requires a balancing role of the State, with the observance of constitutional principles, through laws and counterbalance mechanisms capable of preventing the increase of the concentration of power and exacerbation of existing inequalities and the emergence of new ones (NEMITZ, 2018).

Observing how national AI strategies have treated this relationship between utility and safeguarding rights, Cath and colleagues (2018) warn that there is a mistaken focus when the utility aspect is prioritized and there is no social emphasis on these technologies. The authors

analyzed three reports on AI strategies and regulation published in the USA, the UK and the European Union and found that there is a common feature (which is also recurrent in documents from other countries): the absence of an effective policy that thinks about a "good AI society". The documents studied are not propositional in this sense and are more concerned with applicability issues (such as economics, business etc.) and not with an effective plan to prepare society for a positive synergy with these innovations: "In short, we need a social strategy for AI, not mere tactics" (p. 3).

9 About frontiers of efficiency and control

As previously seen, one of the fundamental aspects of AI systems is the training phase of the algorithms. With some variations, there are two fundamental forms of training: (a) supervised learning and (b) unsupervised learning. The first way involves giving the system data that point exactly to the desired direction, showing the code the desired parameters. Taking our initial example, it would be like giving the algorithm the objective of identifying whether an image is of a bird and, for that, the code learns from a large volume of images of all types of birds so that the system understands which variable patterns are more likely to be successful in identifying birds. That is, we train the algorithm for a specific horizon and offer supervised information to that end.

On the other hand, the unsupervised learning is when we do not give the code specific instructions and let the system randomly consume an amount of information, being able to identify patterns or correlations that are hidden or hardly noticeable to the naked eye. In this model, the algorithm can take different paths and discover things that were not foreseen. Returning to our example, it would be like drawing a code that consumes images of all kinds until it becomes able to perceive patterns and differentiate the images of a bird, an airplane or a butterfly. Especially in the latter case, the unsupervised form, there is great concern about the loss of control of the human element in the system. Once the code is generated and is not properly supervised depending on the area in which the algorithm operates, it may result in harmful results or actions not foreseen by its developers.

Another important dimension that involves both cases is the risk of losing control, but this time not in the final result, but in the path that the machine has chosen to pursue its goal. Russell (2019) argues that one of the main problems of AI is precisely the unmeasured focus on efficiency that, to be achieved, can mean an ethically contestable path:

> As machines designed according to the standard model become more intelligent, however, and as their scope of action becomes more global, the approach becomes untenable. Such machines will pursue their objective, no matter how wrong it is; they will resist attempts to switch them off; and they will acquire any and all resources that contribute to achieving the objective (RUSSELL, 2019, p. 171).

In this sense, the regulatory guidelines and ethical principles in the development of codes have debated the importance of restricting this type of application in more sensitive areas (which affect the lives of individuals more directly and impactful) and, at the same time, creating stricter monitoring mechanisms in the processes capable of making AI systems more accountable, including the possibility of manual shutdown in the system in case of deviation or loss of control.

10 About diversity and representativeness in the codes

A final key issue that has been going through the regulatory and normative debates on Artificial Intelligence concerns the inclusive dimension that the widespread use of algorithms for various everyday applications requires. Specifically, it involves thinking about the values that are embedded (or forgotten) in the code. This brings us to a peculiar form of representativeness in view of the fact that algorithms are representations of values embedded in the machine that, although they seem falsely neutral, are actually the result of a worldview or of objective subjectivities:

> Programmers can create algorithms that have biased assumptions or limitations built into them. They may unconsciously state a question in a biased way. The prejudices of individual programmers can have a wide and cumulative effect because, in a complex software system composed of smaller subsystems, the real bias of the system can be a composite of rules specified by different programmers (CITRON, 2008, p. 1262).

This applies to both supervised and unsupervised machine learning systems because the choice of data to be used by the algorithms in the training phase may be a biased choice or the data itself or the training environment can contain a number of bias problems. Several studies have demonstrated the discriminatory effect that algorithms can reinforce (GRAHAM, 2004; CPL, 2017; LEURS; SHEPHERD, 2017; WEST et al 2019) and this occurs both directly and deliberately as well as indirectly, as a reflection of the systems design productive process, that is, of the level of the programmers' teams:

> Discrimination and inequity in the workplace have significant material consequences, particularly for the under-represented groups who are excluded from resources and opportunities. For this reason alone the diversity crisis in the AI sector needs to be urgently addressed. But in the case of AI, the stakes are higher: these patterns of discrimination and exclusion reverberate well beyond the workplace into the wider world. Industrial AI systems are increasingly playing a role in our social and political institutions, including in education, healthcare, hiring, and criminal justice. Therefore, we need to consider the relationship between the workplace diversity crisis and the problems with bias and discrimination in AI systems (WEST et al, 2019, p. 15).

For this reason, several organizations and specialists have been addressing the problem of representativeness in the process of building systems. This goes through both the existence of deontological codes that programmers must follow to incorporate greater diversity and worldviews - even those that are absent, going beyond their social circle - as well as to break with ethnic, cultural and racial tendencies of those who draw the code and build AI systems that will ultimately be an expanded reflection of human yearnings, for better or worse.

11 Conclusion

The main objective of this article was to synthesize and characterize the main key problems that a good Artificial Intelligence policy needs to answer, in the context of contemporary democracies. Initially, the concern was to synthesize the emergence and evolution of Artificial Intelligence as a field of studies and development, as well as highlighting the main aspects that help us better understand its way of functioning. In this sense, initially, the statistical role of artificial neural networks was discussed, the strength and the problems behind the idea of "imitation", the power of repetition in automation mechanisms, and the types or levels of AI.

Then, from the perspective of the "boom" in government and non-government documents on AI strategies and policies published mainly in 2017, seven important key issues of political emphasis were raised that are at the foundation of this debate: (1) personification and imputability of the machine; (2) dilemmas and moral judgments; (3) statistical authoritarianism of the metrics; (4) mathematical obscurantism in the processes; (5) utilitarian pervasiveness of autonomous systems; (6) frontiers of efficiency and control; (7) diversity and representativeness in the codes.

Far from exhausting all the dimensions that this complex phenomenon raises, the proposal was to highlight a set of problems that have political implications and that deserve special attention because they are configured as determining elements for the good relationship between democracy and the Artificial Intelligence systems on the rise. The idea is to collaborate with a broader and political understanding of this phenomenon, useful to foster the deepening of the debate on regulatory strategies and policies on AI under construction or consolidation, strengthening a more humanistic perspective that goes beyond technical efficiency. Above all, it was observed the dilemmas that this scenario presents, the concentration of power that is insinuated and the possible increase in asymmetries in increasingly datafied societies. This whole scenario requires the State to assume its proper role as protector of individual and collective rights, putting into practice plans, policies and appropriate regulation for this field.

References

ATABEKOV, A.; YASTREBOV, O. Legal Status of Artificial Intelligence Across Countries: Legislation on the Move. **European Research Studies Journal**, v. 21, n. 4, p. 773-782, 2018.

BENTHAM, Jeremy. **An Introduction to the Principles of Morals and Legislation**. Kitchener: Batoche Books, 2000.

BIGONHA, Carolina. Inteligência Artificial em perspectiva. **Panorama Setorial da Internet**, n.2. São Paulo: NIC.Br, 2018, p. 1-9.

BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. **Big Data & Society,** 2016, p. 1-12;

CATH, C. *et al.* Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. **Science and Engineering Ethics**, v. 24, n. 2, 2018, p. 505-528.

CITRON, Danielle Keats. Technological Due Process. **Washington University Law Review**, v. 85, n. 6, p. 1249-1313, 2008.

COPELAND, Jack. Artificial Intelligence. *In:* COPELAND, Jack (org).**The Essential Turing**: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life: Plus The Secrets of Enigma. Oxford: Oxford University Press, 2004, p. 353-261.

CPL. Centre for Public Impact. **Destination unknown:** Exploring the impact of Artificial Intelligence on Government - Working Paper.Londres: CPL. Centre for Public Impact, 2017. Disponível em: https://resources.centreforpublicimpact.org/production/2017/09/Destination-Unknown-AI-and-government.pdf. Acesso em: 15 jul. 2019.

DONEDA et al. Considerações iniciais sobre inteligência artificial, ética e autonomia pessoal. Pensar: **Revista de Ciências Jurídicas**, v. 23, n. 4, p. 1-17, 2018.

GIRASA, Rosario. **Artificial Intelligence as a Disruptive Technology**: Economic Transformation and Government Regulation. Cham: Palgrave Macmillan, 2020

GLEICK, James. **A informação**: Uma história, uma teoria, uma enxurrada. São Paulo: Compahia das Letras, 2011.

GRACE, Katja *et al*. When Will AI Exceed Human Performance? Evidence from AI Experts. **Journal of Artificial Intelligence Research**, n. 62, p. 729-754, 2018.

GRAHAM, Stephen. The woftware-sorted city: rething the "digital divide". *In:* GRAHAM, S. (org.). **The cybercities reader**. Londres: Routledge, 2004, p. 324-332.

GUTMANN, Amy; THOMPSON, Dennis. **Democracy and Disagreement**. London: Cambridge; Massachusetts: Harvard University Press, 1996.

HEBB, Donald O. **The organization of behavior**: a neuropsychological theory. Oxford: Wiley, 1949.

HEIDEGGER, Martin. **Ensaios e conferências**. Petrópolis: Vozes, 2001.

HELBING, Dirk et al. Will Democracy Survive Big Data and Artificial Intelligence? **Scientific American**, n. 98, p.73-98, *2*017. Disponível em: www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence. Acesso em: 15 jul. 2019.

LEURS, Koen; SHEPHERD, Tamara. Datafication & Discrimination. *In:* SCHÄFER, Mirko Tobias; ES, Karin van (org). **The Datafied Society**: Studying Culture through Data. Amsterdam: Amsterdam University Press, 2017, p. 211-231.

McCULLOCH, Warren S.; PITTS, Walter H. A logical calculus of the ideas immanet in nervous activity. **Bulletin of Mathematical Biophysics**, n. 5, p. 115-133, 1943.

LEESE, Matthias. The new profiling: Algorithms, black boxes, and the failure of ant)i-discriminatory safeguards in the European Union. **Security Dialogue**, v. 45, n.5, p. 494-511, 2014.

MAYER-SCHONBERGER, Viktor; CUKIER, Kenneth. **Big Data:** como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana. Rio de Janeiro: Editora Campus, 2013.

MITTELSTADT, B. D. *et al*. The ethics of algorithms: Mapping the debate. **Big Data & Society**, v. 3, n. 2, p. 1-21, 2016

MONTRÉAL DECLARATION, For a **Responsible Development of Artificial Intelligence 2018**. Montreal: Université de Montréal, 2018. Disponível em: https://www.montrealdeclaration-responsibleai.com/the-declaration. Acesso em: 22 jul. 2019.

NEMITZ, Paul. Constitutional democracy and technology in the age of artificial intelligence. **Philosophical Transaction Real Society**, p. 1-14, 2018.

RENDTORFF-SMITH, Sara. Desafios de Governança em Inteligência Artificial. Entrevista. In: **Panorama setorial da Internet**, n.2, São Paulo: NIC.Br, 2018, p. 10-12.

ROSENBLAT, Frank. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychologial Review**, n. 65, p. 386-408, 1958.

RUSSELL, Stuart. **Human Compatible**: artificial intelligence and the problem of control, 2019.

SCHERER, Matthew U. Regulating artificial intelligence systems: reisks, challenges, competencies, and strategie. **Harvard Journal of Law & Technology**, v. 29, n. 2, p. 354-400, 2016.

SILVA, Sivaldo P. da. Comunicação digital, Economia de Dados e a racionalização do tempo: algoritmos, mercado e controle na era dos bits. **Revista Contracampo**, v. 38, n. 1, p. 157-169, 2019.

TURING, A.M. Computing machinery and intelligence**. Mind**: a Quarterly Review of Psychology and Philosophy, v.49, n. 236, p. 433-460, 1950.

TUTT, Andrew. An FDA for Algorithms. **Admin Law Review**, n. 83, p. 83-123, 2017.

WEST, S.M., Whittaker, M. and Crawford, K. **Discriminating Systems:** Gender, Race and Power in AI. Nova York: AI Now Institute. 2019. Disponível em: https://ainowinstitute.org/discriminatingsystems.html. Acesso em: 8 fev. 2020.Article submitted on 2020-05-02

Article submitted on: 2020-05-02

Article accepted on 2020-05-31