Paakat
Revista de tecnología y sociedad

Models and good assessment practices to detect impacts, risks and damages of artificial inteligence

Modelos y buenas prácticas evaluativas para detectarimpactos, riesgos y daños de la inteligencia artificial

http://dx.doi.org/10.32870/Pk.a12n23.742

Jorge Francisco Aguirre Sala*
http://orcid.org/0000-0002-5805-4082
Universidad Autónoma de Nuevo León, México

Received: Febraury 25, 2022
Accepted: June 16, 2022

ABSTRACT

Starting from exemplifying and recognizing the impacts, risks and damages caused by some artificial intelligence systems, and under the argument that the ethics of artificial intelligence and its current legal framework are insufficient, the first objective of this paper is to analyze the models and evaluative practices of algorithmic impacts to astimate which are the most desirable. The second objective is to show what elements algorithmic impact assessments should have. The theoretical basis for the analysis of models, taken from Hacker (2018), starts from showing the discrimination due to lack of guarantees that the input data is representative, complete, and purged of biases, in particular historical bias coming from representations made by intermediaries. The design to discover the most desirable evaluation instrument establishes a screening among models and their respective inclusión of the elements present in the best practices at a global level. The analysis sought to review all algorithmic impact evaluations in the relevant literature at the years 2020 and 2021 to gather the most significant lessons of good evaluation practices. The results show the convenience of focusing on the risk model and six essential elements in evaluations. The conclusions suggest proposals to move towards quantitative expressions of qualitative aspects, while warning of the difficulties in building a standardized evaluation formula. It is proposed to establish four levels: neutral impacts, risks, reversible and irreversible damage, as well as four protection actions: risk prevention, mitigation, repair and prohibition.

Keywords
Algorithmic risks; evaluative approaches; human decisions on artificial intelligence; sectors and domains

RESUMEN

Tomando como punto de partida el ejemplificar y reconocer los impactos, riesgos y daños causados por algunos sistemas de inteligencia artificial, y bajo el argumento de que la ética de la inteligencia artificial y su marco jurídico actual son insuficientes, el primer objetivo de este trabajo es analizar los modelos y prácticas evaluativas de los impactos algorítmicos para estimar cuáles son los más deseables. Como segundo objetivo se busca mostrar qué

Palabras clave
Riesgos algorítmicos; enfoques evaluativos; decisiones humanas sobre la inteligencia artificial; sectores y dominios.

* PhD in Philosophy from the Universidad Iberoamericana in Mexico City. Currently attached to the Universidad Autónoma de Nuevo León where he is the leader of the Democracy and Sustainability academic body. Member of the SNI, level 2. His lines of research are: liquid democracy, electronic democracy and the emergence of contemporary digital technology. Among his recent publications, the following stands out: Aguirre Sala, J. F. (2021). *¿Qué es la democracia? La transición política por la transformación difital de la democracia*. Tirant lo Blanch. [Research ID: AFM-9989-2022].

elementos deben poseer las evaluaciones de impacto algorítmico. La base teórica para el análisis de modelos, tomada de Hacker (2018), parte de mostrar la discriminación por falta de garantías para que los datos de entrada sean representativos, completos y depurados de sesgos, en particular del sesgo histórico proveniente de representaciones hechas por intermediarios. El diseño para descubrir el instrumento de evaluación más deseable establece una criba entre los modelos y su respectiva inclusión de los elementos presentes en las mejores prácticas a nivel global. El análisis procuró revisar todas las evaluaciones de impacto algorítmico en la literatura atingente de los años 2020 y 2021 para recabar las lecciones más significativas de las buenas prácticas de evaluación. Los resultados arrojan la conveniencia de enfocarse en el modelo del riesgo y en seis elementos imprescindibles en las evaluaciones. En las conclusiones se sugieren propuestas para transitar hacia expresiones cuantitativas de los aspectos cualitativos, a la vez que advierten de las dificultades para construir una fórmula estandarizada de evaluación. Se propone establecer cuatro niveles: impactos neutros, riesgos, daños reversibles e irreversibles, así como cuatro acciones de protección: prevención de riesgos, mitigación, reparación y prohibición.

INTRODUCTION

Although artificial intelligence facilitates the handling of big data, the indiscriminate use of algorithms and machine learning has caused damage and undesirable repercussions in the decision-making process. Therefore, discovering the best models and practices to evaluate these algorithmic impacts becomes relevant.

Achieving this objective requires, on the one hand, recognition of the transversality of impacts, risks and damages, and on the other hand, the analysis of models and consideration of the best evaluation practices. With this in mind, the central questions of the inquiry are: what are the most desirable models and practices for assessing the impacts, risks and harms caused by the use of artificial intelligence, and what elements should algorithmic impact assessments possess?

In order to highlight the relevance of artificial intelligence assessments, it is useful to understand two starting points: the nature of artificial intelligence and the various impacts, risks and harms of artificial intelligence. Both aspects will be developed in these introductory paragraphs in order to later provide a guideline for the analysis of the models, the review of the most recognized good practices and, finally, in the conclusions, to reflect on the answers to the research questions, as well as to provide some complementary proposals.

Artificial intelligence is defined by the group of experts of the United Nations Educational, Scientific and Cultural Organization (UNESCO) as: “systems capable of processing data and information in a way that resembles intelligent behavior, and generally encompasses aspects of reasoning, learning, perception, prediction, planning or control” (2021, p. 16).

From this concept it is important to highlight the aspects of learning (by the self-learning of technological systems, machine learning) and control (by direct decision

making or orientations for humans to make decisions). In other words, artificial intelligence algorithms can obtain new knowledge from the basic first-level data or from the layers with which they are fed and, consequently, make or guide decisions faster and with greater probabilistic certainty than human capabilities.

In relation to the algorithms' procedure, it should be noted that the learning and new conclusions obtained by the information technology systems can be supervised or unsupervised. In unsupervised cases, the algorithms obtain the new data from the first results of preliminary inferences, especially from those that were not labeled.

A common example is illustrated in so-called spam, where the email server algorithm labels some emails as "unwanted" based on previous user information or decisions about the sender. Likewise, the algorithm also "decides" that other senders correspond to unwanted mail based on more complex inferences it makes with the initial data. However, making decisions about "unwanted" mail is not as serious as the decision of an algorithmic system that denies entry to a country to a migrant because it did not associate his last name with that of previously qualified foreigners or because it linked him to an offender.

To summarize, artificial intelligence systems receive data, process data under a scheme, program or system and provide an output or information response; however, when the first outputs are processed again by new self-managed schemes from the previous experience, an indeterminate number of hidden layers with increasingly complex representations, correlations and abstractions are produced. It is at this point that deep learning begins.

The process and the new layers that are being created are difficult to trace, to the point that "in deep learning environments, even developers may not be able to 'understand' the reasoning behind some output" (Martinez-Ramill, 2021, p. 4). The complexity and cognitive difficulty of the new layers and their outputs discusses the right of users to the reasonability of the algorithms, while justifying, more solidly, the need for impact assessments.

With regard to the impacts, risks and harms of artificial intelligence, six areas of vulnerability have been identified: risks to citizen security; risks of violations of fundamental rights; lack of procedures and resources on the part of authorities to ensure compliance with regulations; legal uncertainty that deters companies from developing artificial intelligence systems; distrust of artificial intelligence, born of the likely reduction in global competitiveness for companies and governments; and legal inconsistencies between nations that cause obstacles to a cross-border single market and threaten the digital sovereignty of any nation (Dalli, 2021).

In short, vulnerability is found in that what is prohibited in one country is promoted in another, in that what may be mandatory for the authorities of one nation may be considered a crime in another. An example of this can be seen in the prohibition of the Uber platform in Colombia for reasons of free commercial competition, a situation that would not be accepted in the liberal market of the United States of America. A second case is what happened with Corona-Warn-App, which takes data from mobile

devices in order to control covid-19 infection chains; its use was mandatory in China and Korea, while in Germany it is prohibited.

In seeking to categorize the content of vulnerable areas, it is proposed to identify algorithmic effects in impacts, risks and damages, which can be located in any of the six vulnerable areas and in more than one domain. The impacts can be exemplified by situated and non-placed artificial intelligence systems on robots.

A situated case is the ASIMO (Advanced Step in Innovative Mobility) robot created by Honda in 2000 and refined to the 2011 version. The company describes it as "an autonomous machine with the ability to make decisions and make changes in its behavior according to the environment it is in" (Honda-Robotics, n.d.). ASIMO can assist any person in their mobility needs and can be used to replace humans in highly dangerous tasks, such as fighting fires, entering toxic areas or being exposed to war attacks. The particularity of its artificial intelligence consists of responding to environmental stimuli by correcting its trajectory or behavior independently thanks to the coordination of visual and auditory sensors, and it also has the ability to recognize faces and the voice of other people.

A paradigmatic case of the results of non-located artificial intelligence is the Libratus program, which is able to operate successfully in decision making even with incomplete, fraudulently omitted or even misleading information. Its developers, at Carnegie Mellon University, project its performance in decision-making in board games as well as in military strategies, medical treatments, commercial negotiations and, of course, in the field of political decisions in the private and public sector.

In parallel, impacts can become risks. For example, portable digital devices, such as watches with biometric sensors (oximeters, step counters, calorie burners, heart rates, etc.), sports programming devices (such as those of Fitbit and Nike) and GPS trackers, have controlled the lives of most of their consumers regardless of the fact that they can yield false positive and negative results (De Moya & Pallud, 2020; Ruckenstein & Schüll, 2017), which, on a social scale, lead to major errors.

Denying social or medical benefits when there are needs and rights to them or granting them when there is no right to receive them are errors that lacerate the quality of public administration and the rule of law, impoverish social resources and increase inequalities and exclusion. The risks can reach not only the violation of rights, but also cause damages, some that can be mitigated and others that are definitive.

To mention a few, mitigable harms include the interaction of a user with a chatbot, where the algorithm sorts and classifies the user's data and determines or stereotypes his or her condition; while definitive harms include making irreversible decisions based on partial information and procedures, such as classifying a woman exempt from the risk of domestic violence, releasing a criminal with a high probability of recidivism (Hartmann & Wenzelburger, 2021), or categorizing subjects as inappropriate for granting a credit or a scholarship.

Risks and harms present in the decisions yielded by artificial intelligence have a discriminatory basis. According to Hacker (2018), these reasons are due to biased feed data that will produce the inequitable results (pp. 1143-1148). There are few guarantees that the input data in an algorithm is representative, complete and purged of bias, therefore, designers would not be able to claim that the output data is harmonized with ethical principles and legislations –example of this is seen in the exclusion of women in certain job hires in the case of Amazon company (Dastin, 2018).

The inappropriate construction of data sets and tagging is also discriminatory. This has led to, for example, the legal invalidity of smart contracts (Argelich, 2020), or Google Photos facial recognition software mistakenly labeling two people of color as "gorillas" (Zhang, 2015). Another discriminatory reason is due to historical bias in the data and intermediary (proxy) representations.

This can be observed in the variable "race" that has been operationalized in several cities in the United States of America, to the extent that some police practices were denounced for illegal arrests of people of color or with Latino features when artificial intelligence systems were used for facial, voice or gait identification (European Union, Agency for Fundamental Rights, 2020, p. 34).

Other discriminatory cases are associated with the correlation of economic solvency in the case of mortgages, longevity, in terms of health and life insurance, or variables such as "zip-code-sex-age", in relation to car insurance costs. Aizenberg and Van den Hoven (2020) have shown that developers and designers of artificial intelligence systems do not possess a deep understanding of the social and historical reasons for discrimination, as their work concentrates on technical aspects, such as the representativeness of variables and the construction of labels to classify (p. 3). This again shows the need to use algorithmic evaluation models and to turn to good practices.

Models for assessing artificial intelligence

Impacts, risks and harms can be assessed with various models depending on the approaches. In a first set, the focus is on ethics, legality and culture. Some organizations and governments seek to establish codes of ethics on artificial intelligence; the Chinese government has promoted one of the most recent codes in this area (Del Rio, 2022) and even UNESCO (2021) has postulated its recommendations in this regard.

However, these codes are neither binding nor persuasive for all audiences, and the ethics of artificial intelligence (branch of ethics focused on the existence of intelligent robots or any type of artificial intelligence) is only for those who wish to adopt it (Cortina, 2019; Lauer, 2021).

As can be observed, the model of legality is fragmented and has inconsistencies due to the variety of international legislations. This is not limited to what is mandatory in one country and prohibited in another, but also to the dissimilar scope of copyright of AI developers in different jurisdictions. The European Union (European Union,

European Commission, 2021b) has made efforts to reach harmonized legislations, but the authorities are hampered in their judicial intervention by the corresponding copyrights of the developers and owners of the algorithms.

The model of culture is more fragmented than that of legality because of the diverse geolocation of the digital world in Oceania, Asia, Europe, Africa and America. Nevertheless, within digital culture, a leading figure remains Tim Berners Lee (the creator of the World Wide Web), who in November 2019 called for adherence to the so-called Contract for the Web. This action plan consists of nine principles that generate commitments addressed to governments, companies and users, with the aim of keeping the network free, decentralized and secure.

This effort by Berners Lee is a sign of the seriousness of the state of the issue, a concern that the European Union continued to address when establishing the *White Paper on artificial intelligence* in 2020. Other models focus their assessments on collective damage, diagnosis and threat assignment.

From collective harms can be exemplified by the "personalized" product recommendations offered by companies such as Uber, Airbnb, Amazon, Netflix, YouTube and a long etcetera. The use of algorithms in recommendations segments users and causes social fragmentation, eroding community cohesion and solidarity (Yeung, 2019, p. 24).

For their part, diagnostic models can proceed from detections in the system development phases to adjustments by autonomous learning. Unfortunately, these models operate once the damage has been caused (even if they occur in the early development phase).

As an example of this, the driverless cars, created in Germany by Ernst Dickmanns since 1986, have a system capable of inputting light, chromatic, audible, tactile, geopositioned, kinemetric, thermal, etc. data, where the output data can be incompatible with the output of the system, where the output data can be incompatible with the input data, for example, in an imminent collision, the input data on the object to collide can vary the output if it is an inanimate object (hitting a pole or a tree), or an animated object (a deer or a cyclist).

In this case, in its programming the system has responsibility, self-learning and adaptation, characteristics that can lead to conflicts with ideas such as not involving third parties in risks, where it is a possibility that the self-driving system ends up sacrificing several crew members to avoid running over a mouse.

The threat assessment model is partially comparable to that of risk and responsibilities, which will be analyzed below. At this point, it is worth questioning who assigns the threats and to whom they are assigned. There is undoubtedly a dialectic between the interests of user agents (e.g., companies or government institutions that use platforms to perform their services) and the interests of end consumers or citizens, who are more likely to perceive themselves as victims. The allocation of responsibility is debatable when only algorithmic platforms can be acted upon and these are backed by the non-negotiable statements that end in the familiar coercion of "I accept the terms of use".

The Council of Europe, through the study conducted by Yeung (2019), shows four models for evaluating artificial intelligence and its algorithms. The first one is based on intent and guilt and focuses on the identification and legal nature of the operators.

In this model there can be several layers of responsibility: a first layer is occupied by the clients or funders who commission designers and developers to build a system; the second layer corresponds to the latter together with the operators and programmers; a third layer could be assigned to the systems themselves, because of their autonomous self-learning capacity; and the final layers of users and consumers are added.

In contrast to the previous one, the second model is based on risk. This model is preventive; it seeks to avoid negligence by investigating possible risks in users and consumers throughout the life of the system. However, all possible dangers are not foreseeable due to the self-learning capacity and programming autonomy of advanced systems.

Therefore, algorithmic impact assessments with differentiated capability throughout the life of a system, including the phases of errors, experimentation, inputs and outputs with unusual information, as well as the self-learning layer, are necessary.

This risk-based model calls into question the responsibility of the different actors. Funders, designers and developers should be exempted from liability when users use systems for purposes other than those offered or perform actions that go beyond the original intentions. The responsibilities attributed to design cannot be equated with those of self-learning or negligent or ill-intentioned use by the end consumer.

The third model corresponds to legal liability, which arises from deficiencies and defects in the systems. To mention an example, when in the automated decisions of delayed self-learning a system puts in the hands of humans the final decision or execution with delay (the control of a self-directed vehicle, the transfer to a low security prison of a highly dangerous prisoner, etc.).

Legal liability, like any criminalization, can anticipate harm and transgression of rights and, due to the novelty of artificial intelligence applications, needs to be detected with impact assessments. The cases noted above on Amazon's employment discrimination, Google's racial marginalization, the arbitrariness of personalized blockchain contracts or false positives and negatives in many other domains, reiterate the need for algorithmic impact assessments.

The fourth model relates to mandatory insurance and focuses on compensation rather than concentrating on forecasting or prevention. Instituting compulsory insurance at the expense of the end users or consumers of artificial intelligence systems might not always be satisfactory, neither because of the costs of the policies nor because of the amounts of compensation and indemnities.

The economic and power interests that resist the models of legal liability and compulsory insurance are notorious: officially the European Union and the US

government are very aware of the mitigation of costs and regulations to avoid competitive restrictions in order to protect their own leadership.[1]

On the other hand, algorithmic liability dissolves when the corporations that built the artificial intelligence programs disappear or change their corporate identity. To avoid this vacuum, and the impunity created thereby, in some countries they promote to have a responsible person within reach fixed by assigning legal capacity to the artificial intelligence programs located (this in the case of robots) (Henz, 2021). The debate on the legal capacity of artificial intelligence systems has many edges, is unfinished and focuses attention only on the aspect of the legality model.

In the face of the preventive model of risk, proposals focused on establishing liability, finding guilty or negligent agents and having compulsory insurance are less desirable. No matter how generous the compensation for a plane crash due to the fault of the algorithm in the control tower or the aircraft, the avoidable loss of life will in many ways be irreparable.

It can be argued that the risk assessment model is optimal because of its preventive capacity and its comprehensiveness over the life of the systems, as well as the evaluation of self-learning processes and the avoidance of repair costs through self-correction and even prohibition actions. In short, the commercial, legal and technical environments in which artificial intelligence and algorithms emerge have rendered ethical principles such as transparency, explicability, accuracy, auditability, accountability and co-construction inoperative.

Consequently, the case for establishing algorithmic impact assessments to support the foundations of ethical judgments, legal judgments and liability rulings is strengthening. As the responsible director of PricewaterhouseCoopers International Limited (PwCIL) in the United States of America points out, for artificial intelligence "academics, non-governmental organizations (NGOs) and some policy makers recommend the adoption of algorithmic Impact Assessments" (Golbin, 2021).

Best practices of algorithmic impact assessment

Algorithmic impact assessments are far from homologous and uniform. The Artificial Intelligence Observatory of the Organization for Economic Co-operation and Development (OECD), in its mission to "conduct an impact assessment and technology foresight on AI" (OECD.AI, 2019, 2021) notes varied public policies and the heterogeneity of strategies, instruments, standards, guidelines, covenants, codes, approaches, canons of considerations on the application and limitations of artificial intelligence.

[1] Further information can be found in Guidance for Regulation of Artificial Intelligence Applications M-21-06 Memorandum for the heads of executive departments and agencies and Executive Order 13859, on maintaining U.S. leadership in artificial intelligence (Vought, 2020).

According to the various observatories and reviews, those of the governments of Australia, Canada, United States of America, Japan, New Zealand, United Kingdom and Singapore stand out as good practices in public administration and management (European Union, European Commission, 2021a, pp. 33-34; Andrade & Kontschieder, 2021; Ada Lovelace Institute & AI Now Institute and Open Government Partnership, 2021; OECD.AI, 2021).

With the possible disadvantage of marginalizing some meritorious case, Table 1 shows, by country and practice, evaluations that strive to meet ethical, legal and cultural criteria such as transparency, explicability, accuracy, auditability, accountability and co-construction.

Table 1. Best practices of algorithmic impact evaluations

Country	Name of instrument	Method	Level achieved on criterion	Areas or domains
Australia	Automated Decision-Making: Better Practice Guide	Qualitative	**Medium**: human, social and environmental well-being, equity, transparency, explainability	Uses of data in government services
Canada	Directive on Automated Decision Making	Qualitative and Quantitative	**Very high**: transparency, accountability and responsibility	Administrative decisions of government services in all areas
United State of America (California)	California State Bill No. 10	Qualitative	**Medium**: state custody over criminals	Human rights
Japan	AI Utilization Guidelines	Qualitative	**Medium**: humanism, education, privacy, security, equity, transparency, accountability, innovation	Legality in government decisions
New Zealand	Goverment algorithm transparecy and accountability	Qualitative	**High**: transparency and accountability	Privacy and efficient use of data
UK	Draft ID Auditing Framework and Guidelines For AI Procurement	Qualitative	**High**: governance and accountability; precision and safety	Ethics and security

Country	Name of instrument	Method	Level achieved on criterion	Areas or domains
Singapore	Advisory Council on the Ethical Use of AI and Data	Qualitative	**Medium**: governance	Ethics, communication, companies.

Source: developed by the author from Ada Lovelace Institute and AI Now Institute and Open Government Partnership (2021). Expert Group on Architecture for AI Principles to be Practiced (2021), Andrade and Kontschider (2021), OECD, AI (2021).

With the intention of omitting biases, when analyzing the practices in table 1, it can be stated that the Canadian assessment stands out because it covers the largest possible number of domains with a parameterized methodology, considering individual and community rights, health, well-being and economic interests, as well as the sustainability of the ecosystem and the duration and reversibility of impacts (Government of Canada, 2021).

In addition, the Canadian instrument obtains gross impact and mitigation scores, that is, it is an instrument that takes the qualitative opinions of experts and the forum of those involved in each case, while assigning them weighted scores to establish with greater objectivity the level of algorithmic intervention achieved according to the evaluation criteria.

In Mexico, the Ministry of Public Function developed an instrument derived from the Canadian one that considers the following domains: human rights, equity and social welfare, transparency, accountability and obligations (Government of Mexico, 2018). It is noted that the overlapping domains between both instruments correspond to human rights and welfare. This parameterization not only concerns the domains, but also the dimensions upon which impact levels will have to be established. The Mexican case illustrates the dimensions of data use and management, processes, level of autonomy and functionality of the system, socioeconomic scope and government operations.

Following the analysis of the best practices of algorithmic impact assessment, the elements that any assessment should have can be deduced. Firstly, in line with the observations made on assessment models, they must contain the sources of predictability, risk and negligence. Secondly, they must consider the areas or domains and sectors where they have effects.

Metcalf *et al*. (2021) propose the following elements in impact assessments: the sources of legitimacy, the opinions and qualifications of the stakeholders and the forum of those involved, the catalyzing event that triggers the need for the assessment, the temporality of the system, the level of public access, the method, the set of assessors, the impact itself and, of course, the determination of damages and their corresponding compensation.

To these it is necessary to add the level of autonomy (due to the self-learning of the systems), the methodology of data collection (for reasons of possible invasion of private data) and the management of the system inventories (because these may be temporary or permanent and, consequently, may or may not leave traces of liability).

The authors have shown the variety of areas or domains (mainly fiscal, environmental, human rights, data protection and privacy) along with the different degrees of algorithmic advancement and varied geopolitical dispositions (Metcalf *et al*., 2021). Hence, methodical commensuration between different instruments does not by itself emerge in a congruent and aligned manner among all actors due to vocabularies, metrics, ethical criteria and legal canons.

In an attempt to include all the dimensions, the proposal is to give them a value as a multiplying weighting factor, with the intention of approaching an inclusive but differentiated version of algorithmic impact assessments, to establish a correlation of actions or reactions to be taken in correspondence to the levels of impact, risk and damage.

Table 2 shows the scope of each assessment model according to the minimum indispensable elements of the Metcalf *et al*. (2021) enumeration and exposes the superiority of the model based on the risk approach.

Table 2. Scope of each assessment model according to the indispensable elements in algorithmic impact assessments

Elements / Models	Intent and guilt	Risk	Responsibility	Obligatory insurance
Source of legitimacy	X	**X**	X	-
Actors and forums	X	**X**	-	-
Trigger event	-	**X**	X	X
Temporality	-	**X**	X	-
Public Access	X	**X**	-	X
Method		**X**	X	X
Set of evaluators	X	**X**		X
Determination of damage and compensation	X	**X**	X	X
Total	5	**8**	5	5

Conclusions

The risk model was shown to be predictive, preventive and encompassing of pernicious catalyzing events, therefore, it is undeniably linked to the liability model, since, to the extent that it is predictive, it assists the evolution of legal provisions and insurance obligations. It is also noted that no other model achieves all the elements that an algorithmic impact assessment must have. This conclusion explicitly answers the first question at the heart of this text's inquiry.

By collecting best practice experiences and modeling them to the risk approach, the minimum elements of algorithmic impact assessments should consider that:

1) Algorithmic effects have cross-cutting impacts even if the algorithms are applied as sectoral, therefore, they should contemplate the inclusion of any domain (fiscal, environmental, health, labor, mobility, human rights).
2) The inclusion of the domains should be ranked according to the extent of their cross-cutting presence; therefore, the domains should be considered with parameterized weightings.
3) Actors and subjects involved in the opinion forums, the qualifications and the set of expert evaluators would participate in the qualitative ratings of the algorithmic effects (specially to distinguish between an impact and a risk), with full access to complete documentation and with prior training to identify and satisfy the ethical, legal and cultural principles of transparency, explicability, accuracy, auditability, accountability and co-construction.
4) Methodology, like the information, must be explainable and, above all, open to include and adapt weighting scales for the various domains and effects.
5) In line with the preventive model of risk, assessments should be continued throughout the life of the system.
6) Evaluation of the entire life of the system must include self-learning and deep learning, even when these are exercised autonomously and independently of human intervention.

These six elements answer the second question posed at the beginning of this text. It is necessary to bear in mind that, at present, assessments are presented in different domains, systems have different degrees of progress, and there are varied geopolitical dispositions. This heterogeneity of approaches, methodologies and legal frameworks hinders the standardization of assessments.

While diversity hinders generalized construction, the option of homogenizing assessments is debatable, as a plurality of approaches and methods may be appropriate for novel or unimagined sectoral impacts as artificial intelligence becomes present in more domains of reality. Therefore, algorithmic impact assessments require an open and flexible method to ensure compliance with minimum ethical and legal principles and to remain ad hoc with specific cultural trends of end users and consumers.

Proposals

The ultimate meaning or purpose of the assessing models of algorithmic impacts aims at linking proportional actions with their results, in other words, the assessments should not only ascertain the status of algorithmic impacts, but also guide consequent human interventions. Hence, the correlation between four categorizations of possible outcomes in the assessments and the corresponding protective actions should be postulated.

The categories resulting from the assessments are: definitive and non-definitive damage, risks and simple impacts. The proposed correspondence suggests prohibition in the case of definitive damage, repair in the case of non-definitive damage, mitigation in the case of risks, and prevention in the case of impacts.

It is true that there are difficulties in the construction of a single model (Metcalf *et al*., 2021, p. 51) and that dissimilar metrics of impact levels have been established (Germany has a regulation of five levels while Mexico registers the possibility of four, following the Canadian instrument), but the discussions and cultural trends of digitization insist on moving towards quantitative expressions of qualitative aspects (Government of Mexico, 2018, p. 4).

With all this in consideration, it is worth proposing a line for future research with the intention of achieving the standardized quantitative expressions in accordance to the symmetrical protection actions with the levels of damage or risk. The qualitative categorization (starting point for assigning a damage, risk or simple impact), to be carried out by the subjects involved in an assessment, implies a deliberative exercise and, in terms of public policies (because of the algorithms used in government administration), a governance exercise.

Quantitative allocations will require prior weightings between the domains where artificial intelligence has an impact in order to link them to the different categories evaluated, i.e., it must be known or agreed which variables will have more weight than others. For example, to prioritize between education and medical services, differentiated weighting is required when the distance between marginalizing a person from a school scholarship is notoriously less serious than marginalizing access to a vital health service.

Undoubtedly, these considerations require consensus reached deliberatively within the framework of governance. It should be noted that this is not the case with the Canadian instrument mentioned above because, although it gives the result of the level of impact, its score only considers mitigation.

In summary, the qualitative categorizations and quantitative weightings must comply with the deliberation of the stakeholder forum and the governance guidelines to achieve the ethical and legal principles of transparency and explicability enough to determine, as the case may be, the prohibition of an artificial intelligence system, the repair of its reversible damage, the mitigation of its risks or the prevention of its impacts.

Before concluding, a word of warning: in the current era of digital technologies, software and artificial intelligence systems are available for reaching consensus and

reaching agreements in order to comply with deliberative and governance dynamics (e.g. AgoraVoting, Democracy Os, Liquidfeedback, Appgree, Adhocracy, Titanpad, Loomio, among the most widely used), and artificial intelligence systems are also available for impact assessments and decision making in the face of ethical dilemmas.

For this reason, it would be a paradox if those interested in establishing algorithmic evaluations and ethical judgments on artificial intelligence were to uncritically use algorithmic technologies themselves (De Cremer and Kasparov, 2022). Unfortunately, this could happen with tools such as the Canadian one, which are available to any user online. Given this scenario, critical and constant training of humans to evaluate artificial intelligence is advisable.

REFERENCES

Ada Lovelace Institute & AI Now Institute and Open Government Partnership. (2021). *Algorithmic Accountability for the Public Sector*. Ada Lovelace Institute. https://www.opengovpartnership.org/wp-content/uploads/2021/08/algorithmic-accountability-public-sector.pdf

Aizenberg, E. & Van Den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*, *7*(2), 1-14. https://doi.org/10.1177/2053951720949566

Andrade, N. & Kontschieder, V. (2021). AI Impact Assessment: *A Policy Prototyping Experiment*. Open Loop. http://dx.doi.org/10.2139/ssrn.3772500

Argelich, C. (2020). Smart Contracts O Code Is Law. *InDret*, (2), 1-41. https://doi.org/10.31009/InDret.2020.i2.01

Cortina, A. (2019). Ética de la inteligencia artificial. *Anales de la Real Academia de Ciencias Morales y Políticas*, (96), 379-394. https://www.racmyp.es/docs/anales/a96-24.pdf

Dalli, H. (2021). Artificial intelligence act. European Parliament, European Parliamentary Research Service. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/694212/EPRS_BRI(2021)694212_EN.pdf

Dastin, J. (10 de octubre de 2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

De Cremer, D. y Kasparov, G. (2022). The ethical AI—paradox: why better technology needs more and not less human responsibility. *AI and Ethics*, (2), 1-4. https://doi.org/10.1007/s43681-021-00075-y

De Moya J-F. & Pallud, J. (2020). From panopticon to heautopticon: A new form of surveillance introduced by quantified-self practices. *Information System Journal*, *30*(6), 940-976. https://doi.org/10.1111/isj.12284

Del Río, M. (16 de mayo de 2022). China publica código ético para regular la Inteligencia Artificial, ¿qué diría Isaac Asimov? *Emprendedor*.

https://emprendedor.com/china-codigo-etico-regular-inteligencia-artificial-leyes-robotica-isaac-asimov/

European Union, Agency for Fundamental Rights. (2020). Getting the future right. Artificial Intelligence and Fundamental Rights. Publications Office of the European Union. https://doi.org/10.2811/774118

European Union, European Commission. (2021a) *Annexes accompanying the Proposal for a Regulation of the European Parliament and of the Council. Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.* European Commission. https://eur-lex.europa.eu/resource.html?uri=cellar:0694be88-a373-11eb-958501aa75ed71a1.0001.02/DOC_2&format=PDF

European Union, European Commission (2021b). *Commission staff working document impact assessment. Accompanying the 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*. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021SC0084&qid=1619708088989

Expert Group on Architecture for AI Principles to be Practiced. (2021). AI Governance in Japan Ver. 1.0. Ministry of Economy, Trade and Industry. https://www.meti.go.jp/press/2020/01/20210115003/20210115003-3.pdf

Gobierno de México. (2018). Principios y guía de análisis de impacto para el desarrollo y uso de sistemas basados en inteligencia artificial en la administración pública federal. Secretaría de la Función Pública. https://www.gob.mx/cms/uploads/attachment/file/415644/Consolidado_Comentarios_Consulta_IA__1_.pdf

Golbin, I. (28 de octubre de 2021). Algorithmic impact assessments: What are they and why do you need them? *PricewaterhouseCoopers* US. https://www.pwc.com/us/en/tech-effect/ai-analytics/algorithmic-impact-assessments.html

Government of Canada. (2021). Algorithmic Impact Assessment Tool. Government of Canada. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Hacker, P. (2018). Teaching fairness to artificial intelligence: existing and novel strategies against algorithmic discrimination under EU Law. *Common Market Law Review*, *55*(4), 1143-1183. https://ssrn.com/abstract=3164973

Hartmann, K. & Wenzelburger, G. (2021). Uncertainty, risk, and the use of algorithms in policy decisions: a case study on criminal justice in the USA. *Policy Sciences*, *54*, 269-287. https://doi.org/10.1007/s11077-020-09414-y

Henz, P. (2021). Ethical and legal responsibility for Artificial Intelligence. *Discover Artificial Intelligence*, *1*(2), 1-5 https://doi.org/10.1007/s44163-021-00002-4

Honda-Robotics. (s.f). ASIMO. El robot humanoide más avanzado del mundo. Honda. https://www.honda.mx/asimo

Lauer, D. (2021). You cannot have AI ethics without ethics. *AI and Ethics*, *1*(1), 21-25. https://doi.org/10.1007/s43681-020-00013-4

Martínez-Ramil, P. (2021). Is the EU human rights legal framework able to cope with discriminatory AI? IDP. *Revista de internet, derecho y política*, (34), 1-14. https://doi.org/10.7238/idp.v0i34.387481

Metcalf, J.; Moss, E.; Watkins, E. A.; Singh, R. & Elish, M. C. (2021). Assembling Accountability. *Algorithmic Impact Assessment for the Public Interest. & Society Research Institute.* https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf

Organización para la Cooperación y el Desarrollo Económicos (OECD.AI). (2019). OECD AI Policy Observatory. OECD. https://oecd.ai/en/dashboards/policy-initiatives/2019-data-policyInitiatives-24186

Organización para la Cooperación y el Desarrollo Económicos (OECD.AI). (2021). OECD AI Policy Observatory. OECD. https://oecd.ai/en/dashboards

Organización de las Naciones Unidas para la Educación, la Ciencias y la Cultura (UNESCO). (2021). Proyecto de texto de la recomendación sobre la ética de la inteligencia artificial. En *Informe de la Comisión de Ciencias sociales y Humanas* (pp. 13-42). UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000379920_spa

Ruckenstein, M. & Schüll, N. (2017). The Datafication of health. *Annual Review of Anthropology*, *46*(1), 261-278. https://doi.org/10.1146/annurev-anthro-102116-041244

Unión Europea. (2020). *Libro blanco sobre la inteligencia artificial. Un enfoque europeo orientado a la excelencia y la confianza*. Unión Europea. https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

Vought, R. (2020). Guidance for Regulation of Artificial Intelligence Applications. Executive Office of the President, Office of Management and Budget. https://trumpwhitehouse.archives.gov/wp-content/uploads/2020/11/M-21-06.pdf

Yeung, K. (2019). *Responsibility and AI. A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework.* Council of Europe. https://rm.coe.int/responsability-and-ai-en/168097d9c5

Zhang, M. (1 de junio de 2015). Google Photos Tags Two African-Americans As GorillasThrough Facial Recognition Software. *Forbes*. https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=26a7b-5f4713d

HOW TO CITE

Aguirre Sala, J. F. (2022). Modelos y buenas prácticas evaluativas para detectar impactos, riesgos y daños de la inteligencia artificial. *Paakat: Revista de Tecnología y Sociedad*, *12*(23). http://dx.doi.org/10.32870/Pk.a12n23.742