Paakat: Revista de Tecnología y Sociedad
e-ISSN: 2007-3607
Universidad de Guadalajara
Sistema de Universidad Virtual
México
paakat@udgvirtual.udg.mx

Año 12, número 23, septiembre 2022-febrero 2023

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

Models and good evaluative practices to detect impacts, risks and damages of artificial intelligence

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

[Recibido 25/02/2022. Aceptado para su publicación 16/06/2022]
DOI: http://dx.doi.org/10.32870/Pk.a12n23.742

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 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.

Palabras clave

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

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 inclusion 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.

Introducción

A pesar de que la inteligencia artificial facilita el manejo de muchos datos (*big data*), el uso indiscriminado de algoritmos y del aprendizaje autónomo de los programas (*machine learning*) ha causado daños y repercusiones indeseables en la toma de decisiones al respecto de esta. Por ende, descubrir los mejores modelos y prácticas para evaluar estos impactos algorítmicos adquiere relevancia.

Para lograr este objetivo se requiere, por una parte, el reconocimiento de la transversalidad de los impactos, riesgos y daños, y por la otra, analizar los modelos y considerar las mejores prácticas de evaluación. Con esto en consideración, las preguntas eje de la indagatoria son: ¿cuáles son los modelos y

prácticas evaluativas más deseables de los impactos, riesgos y daños que provoca el uso de la inteligencia artificial?, y ¿qué elementos deben poseer las evaluaciones de impacto algorítmico?

Para dar notoriedad a la relevancia de las evaluaciones de la inteligencia artificial es útil comprender dos puntos de partida: la naturaleza y proceder de la inteligencia artificial y los diversos impactos, riesgos y daños de esta. Ambos aspectos se desarrollarán en estos párrafos introductorios para más adelante dar pauta al análisis de los modelos, la revisión de las buenas prácticas más reconocidas y, finalmente, en las conclusiones reflexionar sobre las respuestas a las interrogantes de investigación, además de aportar algunas propuestas complementarias.

La inteligencia artificial es definida por el grupo de expertos de la Organización de las Naciones Unidas para la Educación, la Ciencias y la Cultura (UNESCO) como: "sistemas capaces de procesar datos e información de una manera que se asemeja a un comportamiento inteligente, y abarca generalmente aspectos de razonamiento, aprendizaje, percepción, predicción, planificación o control" (2021, p. 16).

De este concepto es importante destacar los aspectos de aprendizaje (por el autoaprendizaje de los sistemas tecnológicos, *machine learning*) y del control (por la toma de decisiones directas o las orientaciones para que los humanos tomen las decisiones). Es decir, los algoritmos de la inteligencia artificial pueden obtener nuevos conocimientos a partir de los datos básicos de primer nivel o de las capas con que son alimentados y, en consecuencia, tomar u orientar decisiones más rápidas y con mayor certeza probabilística que las capacidades humanas.

En relación al procedimiento de los algoritmos, cabe destacar que el aprendizaje y las nuevas conclusiones que obtienen los sistemas de información tecnológica pueden ser supervisados o no supervisados. En los casos no supervisados, los algoritmos obtienen los nuevos datos a raíz de los primeros resultados de inferencias preliminares, sobre todo de aquellos que no fueron etiquetados.

Un ejemplo común se ilustra en el llamado correo electrónico no deseado o basura (en inglés *spam*), donde el algoritmo del servidor del correo electrónico etiqueta algunos correos como "no deseados" en base a información o decisiones previas del usuario respecto al remitente. Asimismo, el algoritmo también "decide" que otros remitentes corresponden al correo no deseado según inferencias más complejas que realiza con los datos iniciales. Ahora bien, la toma de decisiones sobre un correo "no deseado" no es tan grave como la decisión de un sistema algorítmico que niega el ingreso a un país a un migrante porque no

asoció su apellido al de extranjeros previamente calificados o porque lo vinculó con un criminal.

Recapitulando, los sistemas de inteligencia artificial reciben los datos, los procesan bajo un esquema, programa o sistema y proporcionan una salida o respuesta de información, no obstante, cuando las primeras salidas vuelven a ser procesadas por nuevos esquemas autogestionados desde la experiencia anterior, se producen una cantidad indeterminada de capas ocultas con representaciones, correlaciones y abstracciones cada vez más complejas. Es en este punto que se inicia el aprendizaje profundo (en inglés *deep learning*).

El proceso y las nuevas capas que se van creando son difíciles de rastrear, hasta el punto de que "en entornos de aprendizaje profundo, incluso los desarrolladores podrían no ser capaces de 'comprender' el razonamiento detrás de cierta producción" (Martínez-Ramill, 2021, p. 4). La complejidad y dificultad cognitiva de las nuevas capas y sus salidas problematiza el derecho que tienen los usuarios a la explicabilidad de los algoritmos, a la vez que justifica, con mayor solidez, la necesidad de las evaluaciones de impacto.

En lo atingente a los impactos, riesgos y daños de la inteligencia artificial se han identificado seis áreas de vulnerabilidad: los riesgos de la seguridad ciudadana; los riesgos de violaciones a los derechos fundamentales; la carencia de procedimientos y recursos por parte de las autoridades para garantizar el cumplimiento de las normativas; la incertidumbre legal que disuade a las empresas para desarrollar sistemas de inteligencia artificial; la desconfianza en la inteligencia artificial, nacida de la probable reducción de la competitividad global para empresas y gobiernos; y las incoherencias jurídicas entre naciones que causan obstáculos para un mercado único transfronterizo y amenazan la soberanía digital de cualquier nación (Dalli, 2021).

En resumen, la vulnerabilidad se encuentra en que lo prohibido en un país se promueve en otro, en que aquello que puede ser obligatorio para las autoridades de una nación, en otra puede tenerse como delito. Ejemplo de esto puede observarse en la prohibición de la plataforma de Uber en Colombia por razones de la libre competencia comercial, situación que no se aceptaría en el mercado liberal de Estados Unidos de América. Un segundo caso es lo ocurrido con Corona-Warn-App, que toma datos desde los dispositivos móviles con el objetivo de controlar las cadenas de contagio por covid-19, su uso fue obligatorio en China y Corea, mientras en Alemania está prohibido.

En busca de categorizar el contenido de las áreas vulnerables, se propone identificar los efectos algorítmicos en impactos, riesgos y daños, que pueden ubicarse en cualquiera de las seis áreas vulnerables y en más de un dominio.

Los impactos pueden ejemplificarse con sistemas de inteligencia artificial situados o no situados en robots.

Un caso situado es el robot ASIMO (por sus siglas en inglés Advanced Step in Innovative Mobility) creado por Honda en el año 2000 y perfeccionado hasta la versión 2011. La empresa lo describe como "una máquina autónoma con la capacidad de tomar decisiones y hacer cambios en su comportamiento de acuerdo con el entorno en el que esté" (Honda-Robotics, s/f). ASIMO puede auxiliar a cualquier persona en sus necesidades de movilidad y puede utilizarse para sustituir a los humanos en tareas de alta peligrosidad, como combatir incendios, ingresar a áreas tóxicas o exponerse a ataques bélicos. La particularidad de su inteligencia artificial consiste en responder a estímulos del entorno corrigiendo su trayectoria o conducta de manera independiente gracias a la coordinación de sensores visuales y auditivos, además tiene la capacidad de reconocer rostros y la voz de otras personas.

Un caso paradigmático de los resultados de la inteligencia artificial no situada es el programa *Libratus*, que es capaz de operar exitosamente en la toma de decisiones aún con información incompleta, omitida con dolo y hasta engañosa. Sus desarrolladores, en la Universidad de Carnegie Mellon, proyectan sus rendimientos en la toma de decisiones tanto en juegos de mesa como en estrategias militares, tratamientos médicos, negociaciones comerciales y, por supuesto, en el ámbito de decisiones políticas en el sector privado y público.

De forma paralela, los impactos pueden llegar a ser riesgos. Por ejemplo, los dispositivos digitales portátiles, como relojes con sensores biométricos (oxímetro, cuenta pasos, cuantificadores de ingesta o quema de calorías, frecuencias cardíacas, etcétera), los dispositivos de programación deportiva (como los de Fitbit y Nike) y los rastreadores GPS, han controlado las vidas de la mayoría de sus consumidores sin importar que pueden arrojan resultados con falsos positivos y negativos (De Moya & Pallud, 2020; Ruckenstein & Schüll, 2017) que, en la escala social, conllevan errores mayores.

Negar prestaciones sociales o médicas cuando se tienen necesidades y derechos a las mismas u otorgarlas cuando no hay derecho a recibirlas son errores que laceran la calidad de la administración pública y el Estado de Derecho, pauperizan los recursos sociales y aumentan las desigualdades y la exclusión. Los riesgos pueden llegar no solo a la vulneración de derechos, sino también a causar daños, algunos que pueden ser mitigados y otros que son definitivos.

Por mencionar algunos, entre los daños mitigables se encuentra la interacción de un usuario con un *chatbot*, donde el algoritmo ordena y clasifica los datos del usuario y determina o estereotipa su condición; mientras que entre

los daños definitivos se cuentan la toma de decisiones irreversibles a partir de información y procedimientos parciales, como clasificar a una mujer exenta del riesgo de violencia doméstica, liberar un criminal con alta probabilidad de reincidencia (Hartmann & Wenzelburger, 2021), o categorizar a sujetos como inapropiados para la ampliación de un crédito o una beca.

Los riesgos y daños presentes en las decisiones arrojadas por la inteligencia artificial tienen una base discriminatoria. Según Hacker (2018), estas razones se deben a datos de alimentación sesgados que producirán lo resultados inequitativos (pp. 1143-1148). Existen pocas garantías para que los datos de entrada en un algoritmo sean representativos, completos y depurados de prejuicios, por tanto, los diseñadores no podrían afirmar que los datos de salida estén armonizados con los principios éticos y las legislaciones –ejemplo de esto se ve en la exclusión de mujeres en ciertas contrataciones laborales en el caso de la compañía Amazon (Dastin, 2018).

La construcción inadecuada de conjuntos de datos y etiquetado (*tagging*) también es discriminatoria. Esto ha dado cabida a, por ejemplo, la invalidez jurídica de los contratos inteligentes (*smart contracts*) (Argelich, 2020), o que el software de reconocimiento facial de Google Photos etiquetara por error a dos personas de color como "gorilas" (Zhang, 2015). Otra razón discriminatoria obedece al sesgo histórico de los datos y las representaciones intermediarias (*proxy*).

Esto puede observarse en la variable "raza" que se ha operativizado en varias ciudades de los Estados Unidos de América, al grado que algunas prácticas policiales fueron denunciadas por detenciones ilegales de personas de color o con rasgos latinos cuando los sistemas de inteligencia artificial se utilizaron para la identificación facial, de la voz o de la manera de caminar (European Union, Agency for Fundamental Rights, 2020, p. 34).

Otros casos discriminatorios están asociados a la correlación de la solvencia económica en el caso de hipotecas, a la longevidad, en cuanto a seguros médicos y de vida, o a variables como "código postal-sexo-edad", en relación con los costos de seguros de automóviles. Aizenberg y Van den Hoven (2020) han mostrado que los desarrolladores y diseñadores de sistemas de inteligencia artificial no poseen una comprensión profunda de las razones sociales e históricas de la discriminación, pues su labor se concentra en aspectos técnicos, como la representatividad de variables y la construcción de etiquetas para clasificar (p. 3). Esto vuelve a mostrar la necesidad de utilizar los modelos de evaluación algorítmica y de acudir a las buenas prácticas.

Modelos para evaluar la inteligencia artificial

Los impactos, riesgos y daños pueden evaluarse con diversos modelos dependiendo de los enfoques. En un primer conjunto, el enfoque se concentra en la ética, la legalidad y la cultura. Algunas organizaciones y gobiernos procuran establecer códigos de ética sobre la inteligencia artificial; el gobierno chino ha promovido uno de los códigos más recientes en esta materia (Del Río, 2022) e incluso la UNESCO (2021) ha postulado sus recomendaciones al respecto.

Empero, estos códigos no son vinculantes ni persuasivos para todas las audiencias, y la ética de la inteligencia artificial (rama de la ética enfocada en la existencia de robots inteligentes o cualquier tipo de inteligencia artificial) es solo para quién desee adoptarla (Cortina, 2019; Lauer, 2021).

Como puede observarse, el modelo de la legalidad está fragmentado y posee inconsistencias por la variedad de legislaciones internacionales. Esto no se limita a lo obligatorio en un país y lo prohibido en otro, sino también al disímil alcance de los derechos de autor de los desarrolladores de inteligencia artificial en distintas jurisdicciones. La Unión Europea (European Union, European Commission, 2021b) ha realizado esfuerzos para alcanzar legislaciones armonizadas, pero las autoridades están impedidas en su intervención judicial por los derechos de autor correspondientes a los desarrolladores y propietarios de los algoritmos.

El modelo de la cultura presenta más fragmentación que el de la legalidad por la diversa geolocalización del mundo digital en Oceanía, Asia, Europa, África y América. No obstante, dentro de la cultura digital, una figura líder sigue siendo Tim Berners Lee (el creador de la *World Wide Web*), quien en noviembre de 2019 convocó a adherirse al llamado *Contrato para la Web*. Este plan de acción consta de nueve principios que generan compromisos dirigidos a gobiernos, empresas y usuarios, con el objeto de mantener a la red libre, descentralizada y segura.

Este esfuerzo de Berners Lee es una muestra de la gravedad del estado de la cuestión, preocupación que continuo la Unión Europea al establecer en 2020 el *Libro blanco sobre la inteligencia artificial*. Otros modelos enfocan sus evaluaciones desde los daños colectivos, el diagnóstico y la asignación de amenazas.

Desde los daños colectivos pueden ejemplificarse las recomendaciones "personalizadas" de productos ofrecidos por empresas como Uber, Airbnb, Amazon, Netflix, YouTube y un largo etcétera. El uso de algoritmos en las recomendaciones segmenta a los usuarios y provoca la fragmentación social, erosionando la cohesión comunitaria y la solidaridad (Yeung, 2019, p. 24).

Por su parte, los modelos de diagnóstico pueden proceder desde las detecciones en las fases de desarrollo del sistema hasta en los ajustes por el aprendizaje autónomo. Lamentablemente estos modelos operan una vez que el daño se ha causado (aunque se den en la fase inicial de desarrollo).

Como muestra de esto, los automóviles autodirigidos (*driverless car*) creados en Alemania por Ernst Dickmanns desde 1986, poseen un sistema con capacidad de entrada de datos lumínicos, cromáticos, audibles, táctiles, geoposicionados, cinemométricos, térmicos, etcétera, donde los datos de salida pueden ser incompatibles con los de entrada, por ejemplo, en una colisión inminente, el dato de entrada sobre el objeto a colisionar puede variar la salida si se trata de un objeto inanimado (chocar contra un poste o un árbol), o un objeto animado (un venado o un ciclista).

En este caso, en su programación el sistema tiene responsabilidad, autoaprendizaje y adaptación, características que pueden llevar a conflictos frente a ideas como la de no involucrar en riesgos a terceros ajenos, donde es una posibilidad que la autoconducción termine sacrificando a varios tripulantes por evitar arrollar a un ratón.

El modelo de evaluación por asignación de amenazas es parcialmente homologable al del riesgo y responsabilidades que se analizara adelante. En este punto, cabe cuestionar sobre quiénes asignan las amenazas y a quiénes les son asignadas. Sin duda hay una dialéctica entre los intereses de los agentes usuarios (por ejemplo, las compañías o instituciones gubernamentales que utilizan plataformas para realizar sus servicios) y los intereses de los consumidores finales o ciudadanos, que son más propensos a percibirse como víctimas. La asignación de la responsabilidad es debatible cuando solo se puede actuar con las plataformas algorítmicas y estas se respaldan con las innegociables declaraciones que terminan en la consabida coerción del "acepto los términos de uso".

El Consejo de Europa, a través del estudio realizado por Yeung (2019), muestra cuatro modelos para evaluar la inteligencia artificial y sus algoritmos. El primero se basa en la intención y culpabilidad y se enfoca en la identificación y la personalidad jurídica de los operadores.

En este modelo pueden existir varias capas de responsabilidad: una primera la ocupan los clientes o financiadores que encargan la construcción de un sistema a diseñadores y desarrolladores; la segunda capa corresponde a estos últimos junto con los operadores y los programadores; una tercera capa podría asignarse a los sistemas mismos, por su capacidad autónoma de autoaprendizaje; y se suman las capas finales de los usuarios y los consumidores.

En contraste con el anterior, el segundo modelo se fundamenta en el riesgo. Este resulta preventivo, desea evitar la negligencia al indagar los posibles riesgos en usuarios y consumidores durante toda la vida del sistema. Sin embargo, la totalidad de los posibles peligros no son previsibles debido a la capacidad de autoaprendizaje y autonomía de programación de los sistemas avanzados.

Por lo tanto, se hacen necesarias las evaluaciones de impacto algorítmico con capacidad diferenciada a lo largo de la vida de un sistema, incluyendo las fases de errores, experimentación, entradas y salidas con información inusual, así como la capa de autoaprendizaje.

Este modelo basado en el riesgo pone en cuestión la responsabilidad de los diferentes actores. Financiadores, diseñadores y desarrolladores deberían quedar exentos de responsabilidad cuando los usuarios utilizan los sistemas para propósitos distintos a los ofrecidos o realizan acciones que van más allá de las intenciones originales. No pueden equipararse las responsabilidades adjudicadas al diseño con las del autoaprendizaje o del uso negligente o mal intencionado del consumidor final.

El tercer modelo corresponde a la responsabilidad legal, esta acontece por déficits y defectos de los sistemas. Por mencionar un ejemplo, cuando en las decisiones automatizadas del autoaprendizaje tardío un sistema pone en manos de humanos la decisión o ejecución final con dilación (el control de un vehículo autodirigido, el traslado a una prisión de baja seguridad de un preso de alta peligrosidad, etcétera).

La responsabilidad legal, como cualquier tipificación de delito, puede anticiparse a los daños y la transgresión a los derechos y, debido a la novedad de las aplicaciones de inteligencia artificial, requiere ser detectada con evaluaciones de impacto. Los casos arriba señalados sobre la discriminación laboral de Amazon, la marginación racial de Google, la arbitrariedad de los contratos personalizados con cadenas de bloques (*blockchain*) o los falsos positivos y negativos en muchos otros dominios, reiteran la necesidad de las evaluaciones de impacto algorítmico.

El cuarto modelo corresponde al seguro obligatorio y se enfoca en la compensación en vez de concentrarse en la previsión o la prevención. Instaurar los seguros obligatorios con cargo a los usuarios o consumidores finales de sistemas de inteligencia artificial no siempre podría resultar satisfactorio, ni por los costos de las pólizas ni por los montos de las compensaciones e indemnizaciones.

Los intereses económicos y de poder que presentan resistencias a los modelos de la responsabilidad legal y seguro obligatorio son notorios: oficialmente la Unión Europea y el gobierno de norteamericano tienen muy

presente la mitigación de costos y regulaciones para evitar restricciones de competitividad en aras de proteger sus propios liderazgos.[1]

Por otra parte, la responsabilidad algorítmica se disuelve cuando desaparecen los corporativos que construyeron los programas de inteligencia artificial o cambian su identidad mercantil. Para evitar este vacío, y la impunidad que genera, en algunos países se promueve tener a un responsable al alcance fijado por medio de la asignación de personalidad jurídica a los programas de inteligencia artificial situados (esto en el caso de los robots) (Henz, 2021). El debate sobre la personalidad jurídica de los sistemas de inteligencia artificial posee muchas aristas, está inacabado y centra la atención solo en el aspecto del modelo de la legalidad.

Ante el modelo preventivo del riesgo, las propuestas enfocadas a fincar responsabilidades, hallar agentes culpables o negligentes y contar con seguros obligatorios son menos deseables. Por muy generosa que fuese la indemnización de un accidente aéreo debido a la culpa del algoritmo en la torre de control o en la aeronave, la pérdida evitable de vidas humanas será en muchos sentidos irreparable.

Puede afirmarse que el modelo evaluativo sobre el riesgo resulta óptimo por su capacidad preventiva y su amplitud a lo largo de la vida de los sistemas, así como por la evaluación de los procesos de autoaprendizaje y por evitar costos reparatorios gracias a las acciones de autocorrección e inclusive de prohibición.En síntesis, los entornos comerciales, jurídicos y técnicos en los que emergen la inteligencia artificial y los algoritmos han provocado la inoperancia de principios éticos como la transparencia, la explicabilidad, la exactitud, la auditabilidad, la rendición de cuentas y la co-construcción.

En consecuencia, se fortalecen las razones para establecer evaluaciones de impacto algorítmico que auxilien los fundamentos de los juicios éticos, las sentencias jurídicas y las resoluciones por responsabilidad. Como señala la directora responsable de PricewaterhouseCoopers International Limited (PwCIL) en los Estados Unidos de América, para la inteligencia artificial "académicos, organizaciones no gubernamentales (ONG) y algunos formuladores de políticas recomiendan la adopción de Evaluaciones de Impacto algorítmicas" (Golbin, 2021).

Buenas prácticas de evaluaciones de impacto algorítmico

Las evaluaciones de impacto algorítmico distan de ser homologables y uniformes. El Observatorio de Inteligencia Artificial de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), en su misión de "realizar una evaluación de

impacto y prospectiva tecnológica sobre la IA" (OECD.AI, 2019, 2021) constata variadas políticas públicas y la heterogeneidad de estrategias, instrumentos, normas, guías, pactos, códigos, enfoques, cánones de consideraciones sobre la aplicación y limitaciones de la inteligencia artificial.

Según los diversos observatorios y revisiones, destacan como buenas prácticas en la administración y gestión pública las de los gobiernos de Australia, Canadá, Estados Unidos de América, Japón, Nueva Zelanda, Reino Unido y Singapur (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).

Con la posible desventaja de marginar algún caso meritorio, en la tabla 1 se muestra, por países y prácticas, las evaluaciones que se esfuerzan por cumplir con criterios éticos, jurídicos y culturales como la transparencia, la explicabilidad, la exactitud, la auditabilidad, la rendición de cuentas y la co-construcción.

Con la pretensión de omitir sesgos, al analizar las prácticas de la tabla 1 puede afirmarse que la evaluación canadiense destaca porque abarca el mayor número posible de dominios con una metodología parametrizada, considerando en lo individual y lo comunitario a los derechos, la salud, el bienestar y los intereses económicos, así como la sostenibilidad del ecosistema y la duración y reversibilidad de los impactos (Government of Canada, 2021).

Además, el instrumento canadiense obtiene puntuaciones de impacto bruto y de mitigación, es decir, es un instrumento que toma las opiniones cualitativas de expertos y del foro de involucrados en cada caso, a la vez que les asigna puntuaciones ponderadas para establecer con mayor objetividad el nivel alcanzado de intervención algorítmica según los criterios de evaluación.

En México, la Secretaría de la Función Pública desarrolló un instrumento derivado del canadiense que considera los siguientes dominios: derechos humanos, equidad y bienestar social, transparencia, responsabilidad y obligaciones (Gobierno de México, 2018). Se observa que los dominios coincidentes entre ambos instrumentos corresponden a los derechos humanos y al bienestar. Esta parametrización no solo concierne a los dominios, sino también a las dimensiones sobre las cuales habrá que establecer los niveles de impacto. El caso mexicano ilustra las dimensiones de uso y gestión de datos, procesos, nivel de autonomía y funcionalidad del sistema, alcance socioeconómico y de las operaciones del gobierno.

Tabla 1. Buenas prácticas de evaluaciones de impacto algorítmico

País	Nombre del instrumento	Método	Nivel alcanzado sobre el criterio	Áreas o dominios
Australia	Automated Decision-Making: Better Practice Guide	Cualitativo	**Medio:** bienestar humano, social y ambiental, equidad, transparencia, explicabilidad	Uso de datos en servicios gubernamentales
Canadá	Directive on Automated Decision Making	Cualitativo y Cuantitativo	**Muy alto:** transparencia, rendición de cuentas y responsabilidad	Decisiones administrativas de servicios gubernamentales en todas las áreas
Estados Unidos de América (California)	California State Bill No. 10	Cualitativo	**Medio:** custodia del Estado sobre criminales	Derechos humanos
Japón	AI Utilization Guidelines	Cualitativo	**Medio:** humanismo, educación, privacidad, seguridad, equidad, transparencia, rendición de cuentas, innovación	Legalidad en decisiones gubernamentales
Nueva Zelanda	Government algorithm transparency and accountability	Cualitativo	**Alto:** transparencia y rendición cuentas	Privacidad y uso eficiente de datos
Reino Unido	Draft AI Auditing Framework and Guidelines For AI Procurement	Cualitativo	**Alto:** gobernanza y rendición de cuentas; precisión y seguridad	Ética y seguridad
Singapur	Advisory Council on the Ethical Use of AI and Data	Cualitativo	**Medio:** gobernanza	Ética, comunicación, empresas

Fuente: elaboración propia a partir de Ada Lovelace Institute & AI Now Institute and Open Government Partnership (2021), Expert Group on Architecture for AI Principles to be Practiced (2021), Andrade & Kontschieder (2021), OECD.AI (2021).

Tras el análisis de las mejores prácticas de evaluación de impacto algorítmico se deducen los elementos que toda evaluación debería poseer. En primer lugar, en congruencia con las observaciones realizadas a los modelos de evaluación, deben contener las fuentes de previsibilidad, riesgo y negligencia. En segundo lugar, han de considerar los ámbitos o dominios y sectores donde tienen efectos.

Metcalf *et al.* (2021) proponen los siguientes elementos en las evaluaciones de impacto: las fuentes de legitimidad, las opiniones y cualificaciones de los actores y del foro de involucrados, el evento catalizador que detona la necesidad de la evaluación, la temporalidad del sistema, el nivel de acceso público, el método, el conjunto de evaluadores, el impacto mismo y, por supuesto, la determinación de daños y su correspondiente compensación.

A estos es necesario añadir el nivel de autonomía (debido al autoaprendizaje de los sistemas), la metodología de recopilación de datos (por razones de la posible invasión a los datos privados) y la administración de los inventarios del sistema (porque estos pueden ser temporales o permanentes y, en consecuencia, dejar o no vestigios de responsabilidad).

Los autores han mostrado la variedad de áreas o dominios (principalmente el fiscal, el medioambiental, los derechos humanos, la protección de datos y la privacidad) junto con los diferentes grados de avance algorítmico y las variadas disposiciones geopolíticas (Metcalf *et al*., 2021). Por ende, la conmensuración metódica entre los distintos instrumentos no emerge por sí misma de manera congruente y alineada entre todos los actores debido a los vocabularios, las métricas, los criterios éticos y los cánones legales.

En un intento de incluir todas las dimensiones, se propone otorgarles un valor como factor multiplicador de ponderación, con la intención de acercarse a una versión incluyente pero diferenciada de las evaluaciones de impacto algorítmico, para establecer una correlación de acciones o reacciones a tomar en correspondencia a los niveles de impacto, riesgo y daño.

En la tabla 2 se muestra el alcance de cada modelo de evaluación según los elementos mínimos indispensables de la enumeración de Metcalf *et al.* (2021) y expone la superioridad del modelo basado en el enfoque de riesgo.

Tabla 2. Alcance de cada modelo de evaluación según los elementos indispensables en las evaluaciones de impacto algorítmico

Elementos / Modelos	Intención y culpabilidad	Riesgo	Responsabilidad legal	Seguro obligatorio
Fuente de legitimidad	x	**x**	x	-
Actores y foros	x	**x**	-	-
Evento detonante	-	**x**	x	x
Temporalidad	-	**x**	x	-
Acceso público	x	**x**	-	x
Método		**x**	x	x
Conjunto de evaluadores	x	**x**		x
Determinación de daño y compensación	x	**x**	x	x
Total	5	**8**	5	5

Fuente: elaboración propia.

Conclusiones

El modelo del riesgo se mostró con capacidad previsoria, preventiva y abarcadora de eventos catalizadores perniciosos, por consiguiente, es innegable que está vinculado al de la responsabilidad, pues en la medida que resulta previsorio auxilia a la evolución de las disposiciones jurídicas y de las obligaciones de seguros. También se observa que ningún otro modelo alcanza todos los elementos que debe poseer una evaluación de impacto algorítmico. Con esta conclusión queda explícitamente respondida la primera pregunta eje de la indagatoria de este texto.

Al recabar las experiencias de las mejores prácticas y modelarlas al enfoque del riesgo, los elementos mínimos de las evaluaciones de impacto algorítmico deben considerar que:

1) Los efectos algorítmicos tienen impactos transversales aunque los algoritmos se apliquen como sectoriales, por lo tanto, deben contemplar la inclusión de cualquier dominio (fiscal, medioambiental, salud, laboral, movilidad, derechos humanos).
2) La inclusión de los dominios debe ser jerarquizada en función de la extensión de su presencia transversal, por esto, los dominios deben considerarse con ponderaciones parametrizadas.
3) Los actores y sujetos involucrados en los foros de opiniones, las cualificaciones y el conjunto de evaluadores expertos participarían en las

calificaciones cualitativas de los efectos algorítmicos (sobre todo para distinguir entre un impacto y un riesgo), con total acceso a la documentación completa y con la capacitación previa para identificar y satisfacer los principios éticos, jurídicos y culturales de transparencia, explicabilidad, exactitud, auditabilidad, rendición de cuentas y co-construcción.

4) La metodología, al igual que la información, debe ser explicable y, sobre todo, abierta a incluir y adaptar las escalas de ponderación de los diversos dominios y efectos.
5) En consonancia con el modelo preventivo del riesgo, las evaluaciones deben continuarse a lo largo de toda la vida del sistema.
6) La evaluación de toda la vida del sistema ha de incluir el autoaprendizaje y aprendizaje profundo, aun cuando estos se ejerzan con autonomía y al margen de la intervención humana.

Estos seis elementos responden la segunda pregunta planteada al inicio de este texto. Es necesario tener en cuenta que, en la actualidad, las evaluaciones se presentan en distintos dominios, los sistemas tienen diferentes grados de avance, y existen variadas disposiciones geopolíticas. Esta heterogeneidad de enfoques, metodologías y marcos legales impide la estandarización de las evaluaciones.

Si bien la diversidad obstaculiza la construcción generalizada, es debatible la opción de homogeneizar las evaluaciones, pues quizá la pluralidad de enfoques y métodos resulte adecuada para impactos sectoriales novedosos o no imaginados gracias a la vertiginosidad con que la inteligencia artificial se hace presente en más dominios de la realidad. Por ende, las evaluaciones de impacto algorítmico requieren de un método abierto y dúctil para garantizar el cumplimiento de principios éticos y jurídicos mínimos y mantenerse *ad hoc* con tendencias culturales específicas de los usuarios y consumidores finales.

Propuestas

El sentido o finalidad última de los modelos evaluativos de los impactos algorítmicos apunta a vincular acciones proporcionales con sus resultados, en otras palabras, las evaluaciones no solo deben constatar el estado de impactos algorítmicos, sino orientar las intervenciones humanas consecuentes. De ahí que deba postularse la correlación entre cuatro categorizaciones de posibles resultados en las evaluaciones y las correspondientes acciones de protección.

Las categorías resultado de las evaluaciones son: daños definitivos y no definitivos, riesgos y simples impactos. La correspondencia propuesta sugiere ante

los daños definitivos la prohibición, ante los no definitivos la reparación, ante los riesgos la mitigación y frente a los impactos la prevención.

Es cierto que la construcción de un modelo único presenta dificultades (Metcalf *et al*., 2021, p. 51) y que se han establecido métricas disimiles de niveles de impacto (Alemania tiene una regulación de cinco niveles mientras México registra la posibilidad de cuatro, siguiendo el instrumento canadiense), pero las discusiones y las tendencias culturales de la digitalización insisten en transitar hacia expresiones cuantitativas de los aspectos cualitativos (Gobierno de México, 2018, p. 4).

Con todo esto en consideración, cabe proponer una línea para futuras investigaciones con la intención de lograr las expresiones cuantitativas estandarizadas de conformidad a las acciones de protección simétricas con los niveles de daño o riesgo. La categorización cualitativa (punto de partida para asignar un daño, riesgo o simple impacto), que deberán llevar a cabo los sujetos involucrados en una evaluación, implica un ejercicio deliberativo y, en términos de políticas públicas (por los algoritmos utilizados en la administración gubernamental), un ejercicio de gobernanza.

Por su parte, las asignaciones cuantitativas requerirán de las previas ponderaciones entre los dominios donde incide la inteligencia artificial para vincularlas con las distintas categorizaciones evaluadas, es decir, debe saberse o consensuarse qué variables tendrán más peso que otras. Por ejemplo, para priorizar entre la educación y los servicios médicos se requiere de la ponderación diferenciada cuando la distancia entre marginar a una persona de una beca escolar es notoriamente menos grave que marginar del acceso a un servicio de salud vital.

Sin duda estas consideraciones requieren de consensos alcanzados deliberativamente en el marco de la gobernanza. Debe advertirse que no ocurre así con el citado instrumento canadiense pues, si bien arroja el resultado del nivel de impacto, su escore solo contempla la mitigación.

En resumen, las categorizaciones cualitativas y las ponderaciones cuantitativas deben cumplir con la deliberación del foro de involucrados y las directrices de gobernanza para alcanzar los principios éticos y legales de transparencia y explicabilidad suficientes para determinar, según sea el caso, la prohibición de un sistema de inteligencia artificial, la reparación de sus daños reversibles, la mitigación de sus riesgos o la prevención de sus impactos.

Antes de finalizar cabe una advertencia: en la actual era de las tecnologías digitales se ofrecen softwares y sistemas de inteligencia artificial para alcanzar consensos y llegar a acuerdos con el propósito de cumplir con dinámicas

deliberativas y de gobernanza (por ejemplo *AgoraVoting*, *Democracy Os*, *Liquidfeedback*, *Appgree*, *Adhocracy*, *Titanpad*, *Loomio*, entre los más utilizados), asimismo existen sistemas de inteligencia artificial disponibles para las evaluaciones de impacto y la toma de decisiones ante dilemas éticos.

Por esta razón sería una paradoja que los interesados en establecer las evaluaciones algorítmicas y los juicios éticos sobre la inteligencia artificial utilizaran de manera acrítica, a su vez, tecnologías algorítmicas (De Cremer y Kasparov, 2022). Lamentablemente, esto podría ocurrir con instrumentos como el canadiense, que están a disposición de cualquier usuario en línea. Ante este escenario, es recomendable la capacitación crítica y constante de los humanos para evaluar la inteligencia artificial.

Referencias

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 Experimen*t. 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 harmonised 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-9585-01aa75ed71a1.0001.02/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 Gorillas Through 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

CÓMO CITAR ESTE ARTÍCULO:

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

* Doctor en Filosofía por la Universidad Iberoamericana de la Ciudad de México. Actualmente adscrito en la Universidad Autónoma de Nuevo León donde es líder del cuerpo académico Democracia y Sustentabilidad. Miembro del SNI, nivel 2. Sus líneas de investigación son: democracia líquida, democracia electrónica y la eclosión de la tecnología digital contemporánea. Entre sus publicaciones recientes destaca: Aguirre Sala, J. F. (2021). *¿Qué es la democracia electrónica? La transición política por la transformación digital de la democracia*. Tirant lo Blanch. [Research ID: AFM-9989-2022].

[1] Para más información, puede consultarse *Guidance for Regulation of Artificial Intelligence Applications M-21-06 Memorandum for the heads of executive departments and agencies* y la Orden Ejecutiva 13859, sobre mantener el liderazgo estadounidense en inteligencia artificial (Vought, 2020).