Inteligencia artificial y derechos humanos: aportes desde la teoría al caso ROMA

Artificial intelligence and human rights: contributions from theory to the ROMA case

Juan Esteban Yanguas Ramón*
Isabella Narváez Peña**

Artículo de reflexión
Fecha de recepción: 31 de agosto de 2024
Fecha de aceptación: 10 de noviembre de 2024

Para citar este artículo:

Yanguas Ramón, J. E. y Narváez Peña, I. (2025). Inteligencia artificial y derechos humanos: aportes desde la teoría al caso ROMA. *Revista Análisis Jurídico-Político, 7*(13), 65-96. https://doi.org/10.22490/26655489.8466

Resumen

El presente artículo examina la obligación constitucional y convencional de las entidades estatales de respetar y proteger los derechos humanos, especialmente en el contexto del diseño de algoritmos

* Abogado por la Universidad Libre de Colombia, seccional Bogotá; magíster en Derecho Constitucional por la Universidad Libre de Colombia, seccional Cali; especialista en Docencia Universitaria por la Universidad Piloto de Colombia y estudiante de la especialización en Ciencias Penales y Criminológicas de la Universidad Externado de Colombia. Coautor de distintos libros enfocados en derecho constitucional, derecho internacional de los derechos humanos, derecho penal e inteligencia artificial. Conferencista invitado en distintos congresos y universidades, tanto nacionales, como internacionales. Actualmente trabaja en la Fiscalía Delegada ante Corte Suprema de Justicia de la Fiscalía General de la Nación. @desenredemoselderecho en redes sociales. Correo electrónico: juanesyanguas@gmail.com ORCID: 0009-0001-9328-1483

** Abogada y especialista en Derecho Administrativo por la Universidad Santiago de Cali, estudiante de la especialización en Derecho Constitucional de la Universidad Externado de Colombia. Actualmente trabaja en la Rama Judicial. @desenredemoselderecho en redes sociales. Correo electrónico: isabellanarvaezpena@gmail.com ORCID: 0009-0009-9770-1690

impulsados por Inteligencia Artificial (IA). El artículo explora, desde la teoría, diversos conceptos, pasando por el *machine learning* y la analítica de datos, así como los derechos humanos desde definiciones informales hasta formales, analizando sus elementos y los sistemas de protección de estos derechos. Se centra en el Sistema Interamericano y el control de convencionalidad, destacando la obligación estatal de integrar un enfoque de derechos humanos en el diseño y supervisión de sus acciones, incluyendo el desarrollo de IA, y la necesidad de consolidar un marco ético basado en derechos humanos para la creación de IA. Finalmente, se analiza, desde una perspectiva constitucional y convencional, el caso del sistema ROMA, creado por la Fiscalía General de la Nación de Colombia, diseñado para identificar factores de reincidencia delictiva, dejando algunas recomendaciones a manera de conclusión.

Palabras clave: Control de Convencionalidad, Derecho Constitucional, Derecho Penal, ROMA.

Abstract

The article examines the constitutional and conventional obligation of state entities to respect and protect human rights, especially in the context of designing algorithms driven by Artificial Intelligence. The article explores various concepts from a theoretical perspective, including machine learning, data analytics, and human rights, from informal to formal definitions, analyzing their elements and the systems for protecting these rights. It focuses on the Inter-American System and the control of conventionality, highlighting the state's obligation to integrate a human rights approach in the design and supervision of its actions, including the development of AI, and the need to establish an ethical framework based on human rights for the creation of AI. Finally, it analyzes, from a constitutional and conventional perspective, the case of the ROMA system, created by the General Attorney's Office of Colombia, designed to identify factors of criminal recidivism, and offers some recommendations as conclusions.

Keywords: Constitutional Law, Conventionality Control, Criminal Law, ROMA.

1. Introducción

La disrupción tecnológica que ha experimentado la humanidad a partir del desarrollo, cada vez mayor, de tecnologías de inteligencia artificial —en adelante, IA— ha generado un amplísimo debate. Para algunos, las tecnologías impulsadas por IA ponen en riesgo la supervivencia de la civilización humana (Harari, 2023). Otros, por el contrario, se muestran optimistas y exponen con frecuencia las bondades de la IA (Gates, 2023).

Lo cierto es que uno de los principales riesgos que plantea la IA, al menos para la comunidad jurídica, es que su crecimiento e implementación se produzcan sin una regulación normativa adecuada, lo cual podría conducir a la vulneración de derechos humanos y al desarrollo de tecnologías de IA que privilegien intereses privados, sin tener en cuenta el bien común (Lara Gálvez, 2020). Construir ese marco normativo sobre bases sólidas es fundamental para garantizar una protección más amplia de los derechos humanos.

Por ello, el presente artículo busca aproximar al lector conceptualmente a la inteligencia artificial, i) exponiendo algunos casos en los que se han denunciado sesgos algorítmicos; ii) se conceptualizará, desde una perspectiva formal, frente a los derechos humanos y sus elementos, para arribar al diseño basado en derechos humanos; iii) se expondrá el caso de ROMA, sistema creado por la Fiscalía General de la Nación de Colombia, para identificar factores de reincidencia y reiteración delictiva; iv) para pasar a hacer un análisis de constitucionalidad y convencionalidad del caso citado.

Finalmente, se elaborarán algunas recomendaciones, a manera de conclusión, para garantizar, no solo que ROMA se ajuste a los parámetros constitucionales y convencionales, sino cualquier otro algoritmo impulsado por IA.

2. Metodología o pauta de análisis

Para abordar lo anterior, se realizó una revisión bibliográfica con el objetivo estructurar una investigación dogmático-jurídica. En ese orden de ideas, a partir de lo establecido teóricamente en la literatura

reciente en materia de inteligencia artificial y derechos humanos, se realizará un análisis aplicado a un caso específico, ROMA, desde los parámetros teóricos del derecho constitucional, del derecho penal y del derecho internacional de los derechos humanos.

3. Desarrollo o núcleo principal y resultados

3.1. Inteligencia artificial para abogados

Cada día se escucha hablar, con más frecuencia, de Inteligencia Artificial, pero ¿qué es? ¿para qué sirve? ¿por qué existe un temor generalizado frente a su implementación? Por ello, el presente subcapítulo desarrollará, en un lenguaje comprensible, el concepto de IA, y algunos casos icónicos en donde se ha señalado a algoritmos de IA de sesgos.

3.1.1. ¿Qué es la IA?

La IA se deriva de la ingeniería informática, y su origen se remonta a la década de los 1950, cuando, durante una conferencia sobre informática en Dartmouth College (Estados Unidos), científicos como John McCarthy, Marvin Minsky, Allen Newell y Hebert Simon presentaron el teorema Logic Theorist, un programador que simulaba las características del ser humano (Benítez *et al.*, 2014).

Otros autores han indicado que los primeros avances de la IA se remontan al siglo XX, entre los años 40 y 50, finalizada la Segunda Guerra Mundial, cuando Alan Turing, uno de los padres de la computación, demostró cómo a través de algoritmos se podía programar los equipos de computación para resolver operaciones básicas (García Serrano, 2012).

La IA es un motor de diálogos entre máquinas y humanos, que imita funciones humanas mediante una programación neurolingüística con base en un soporte de código fuente (informático). Sin embargo, los científicos y teóricos no han podido llegar a un consenso sobre la definición de IA. Lo anterior obedece a que, i) no existe una aceptación generalizada frente a que una máquina pueda incorporar las capacidades mentales propias de los humanos, y, ii) no han podido

definir qué es la inteligencia en los seres humanos (Amador Hidalgo, 1996). Frente a las anteriores afirmaciones surgen los siguientes interrogantes: ¿qué es la inteligencia? ¿qué es la inteligencia humana y qué la diferencia de la IA? ¿puede una máquina incorporar las capacidades mentales propias de los humanos?

El concepto *inteligencia* fue utilizado por primera vez por Francis Galton (1869), quien la definió como una capacidad física de carácter hereditario. A partir de ese momento, se ha suscitado un intenso debate académico interdisciplinario, entre otras, inconcluso, frente a si la inteligencia es innata y heredable, tal como la concebía Galton, o si, por el contrario, depende del entorno, pudiendo mejorarse a partir de los ambientes educativos, culturales, sociales, etc.

Independientemente de su origen, y para efectos del presente artículo, la inteligencia humana es el "conjunto de habilidades cognitivas y conductuales que permite la adaptación eficiente al ambiente físico y social. Incluye la capacidad de resolver problemas, planear, pensar de manera abstracta, comprender ideas complejas, aprender de la experiencia" (Ardila, 2011).

Por otro lado, la IA es una disciplina derivada de la informática, donde un sistema de cómputo a través de su *software* puede realizar actividades emulando los procesos de aprendizaje propios del ser humano. La IA es, entonces, esa función que desarrollan las máquinas, a partir de conocimientos humanos, para realizar sus tareas, y que, si la persona lo realizara, se consideraría inteligente (Romero *et al.*, 2007).

La IA, además de emular los comportamientos inteligentes del ser humano, busca proponer la creación de elementos artificiales. Ello significa que tiene la capacidad de ser entrenada para responder a las demandas de su entorno (Romero *et al.*, 2007). Así, la IA, a través de sus procesos matemáticos y/o algorítmicos, simula los procesos neuronales y comportamientos del cerebro humano, para así dar paso al comportamiento inteligente.

A partir de ello, los proyectos impulsados por IA buscan realizar las mismas tareas que desarrolla el ser humano, en mayor volumen y menor tiempo, no necesariamente para remplazarlos, pero sí para hacer las tareas mejor y más rápido, con un menor margen de error.

Dentro de esas tareas se encuentra la resolución de problemas, uno de los elementos referidos como esenciales de la *inteligencia*. Tras décadas de ensayo y error, el sector privado, principal interesado y promotor de los sistemas de IA, a través de diversas técnicas informáticas logró programar computadoras que pudieran programar otras máquinas programadoras. Así surgió, en la necesidad de automatización, que las máquinas pudieran recrear procesos de autoaprendizaje. Es así como la IA tiene la capacidad, no solo de ejecutar y resolver un "problema", sino de aprender de él, lo que requiere una altísima capacidad de memoria.

Para algunos, la IA supera a la inteligencia humana. Una computadora, Deep Blue, desarrollada por IBM, venció al mejor ajedrecista de la historia, Garry Kaspárov, hace 24 años, en 1996. Desde ese momento, la inteligencia humana dejó de ser competencia para la IA, por lo que hoy día, las verdaderas competencias son entre inteligencias artificiales. AlphaZero, diseñada por la empresa de Google, DeepMind, venció otras máquinas entrenadas para jugar ajedrez, como shogi y go. Sus creadores afirmaron, tras las victorias, que AlphaZero jugó a un nivel "súper humano" (British Broadcasting Corporation [BBC], 2018).

Podría pensarse entonces que todo lo que puede hacer el hombre, la IA podría hacerlo mejor, usada adecuadamente. Sin embargo, el presente artículo se enfocará en la IA como herramienta. En el ejercicio práctico de dicha herramienta, se ha evidenciado que los programas de IA reproducen los sesgos de sus creadores, y como el humano jamás podrá ser objetivo ni imparcial, ya que, desde sus experiencias de vida y su entorno, está contaminado, tampoco la IA como creación de humanos.

3.1.2. ¿Por qué la IA representa un riesgo para el derecho? Análisis de IA sesgada

De manera preliminar, es necesario establecer que los sesgos algorítmicos surgen en aquellos "sistemas cuyas predicciones benefician sistemáticamente a un grupo de individuos frente a otro, resultando así injustas o desiguales" (Ferrante, 2021). Para efectos del presente artículo, la injusticia o desigualdad de aquellas predicciones serán aquellas que

contraríen los derechos humanos. En consecuencia, se entenderán los sesgos algorítmicos como aquellas predicciones que vulneran de manera sistemática derechos humanos. Es evidente que dichos sesgos en los programas impulsados por IA existen y que su surgimiento, en la mayoría de los casos, no suelen ser previstos o premeditados. Lo anterior, implica que, en ocasiones, aunque el programador no tenga la intención de incluir sesgos, la IA igual los aprende.

La inteligencia artificial se está vinculando en áreas cada vez más sensibles, como el sistema de justicia, la contratación en las empresas o entidades del Estado, el sistema de salud, entre otras. Ese crecimiento surgió de la necesidad de agilizar procesos, sin embargo, ha sido cuestionado, por los sesgos que han sido percibidos en la implementación de algunos sistemas de IA, con repercusiones sumamente profundas en la vida y derechos de las personas (Silberg y Manyika, 2019).

En el área de la salud, en la Facultad de Medicina de St. George, Reino Unido, en 1986, se implementó un programa computacional creado para reducir el trabajo de seleccionar candidatos para la entrevista. Buscaba, de manera objetiva, seleccionar a los aplicantes. Sin embargo, algunos candidatos fueron rechazados por su sexo y origen racial, inclusive, denuncian que fueron discriminadas personas con nombres no europeos. Tras un estudio, se identificó que el programa no creó los sesgos, sino que replicó aquellos preexistentes (Lowry y MacPherson, 1988).

La IA también ha sido usada en el área de contratación. Amazon creó un sistema diseñado para agilizar el proceso de reclutamiento mediante la lectura de los CV y la selección del candidato mejor calificado. Literalmente, el hombre mejor calificado. El sistema aprendió por sí mismo que los candidatos masculinos eran preferibles sobre las candidatas femeninas, tras replicar las prácticas de contratación existentes que por lo general presentan sesgos en contra de las mujeres (Dastin, 2018).

En el mundo del mercadeo y las ventas, la minería de datos y, recientemente, los programas impulsados por IA han sido usados para delimitar el público y posibles compradores. En Facebook, los vendedores pueden optar por no anunciar a usuarios de ciertos

vecindarios o grupos étnicos. Así, en lugar de limitar el producto a quienes de verdad lo necesitan excluyen por razones personales a personas de cierto grupo social (MIT Technology Review, 2019).

El Departamento de Vivienda y Desarrollo Urbano (HUD) de Estados Unidos denunció que Facebook extrae los datos de los usuarios y luego los usa para determinar quién ve los anuncios, incluyendo datos como raza, color, religión y sexo. Pero no solo identificaron que pueden ser excluidos de productos a la venta, sino también de ofertas laborales o bienes inmuebles (MIT Technology Review, 2019).

Por ejemplo, los anuncios de búsqueda de profesionales de educación infantil y secretaría se mostraron principalmente a mujeres, mientras que las publicaciones para empleo de conserje y taxista se mostraron en mayor proporción a personas de grupos minoritarios. Los anuncios sobre la venta de las viviendas también aparecieron más a usuarios blancos, mientras que los anuncios de alquileres fueron mostrados más a las minorías (Logically, 2019).

3.2. Derechos humanos, también para abogados

En las noticias se suelen ver titulares frente al riesgo que representa la IA para los derechos humanos. Pero ¿qué son? ¿cuáles son sus características? Este apartado aproximará al lector a una concepción formal de los derechos humanos, a las garantías supralegales de protección de derechos humanos, para aterrizar en el diseño basado en derechos humanos de algoritmos impulsados por IA.

3.2.1. ¿Qué son los derechos humanos?

Los derechos han sido i) privilegios concedidos a grupos poblaciones específicos; ii) precedidos por conflictos o revoluciones; iii) consagrados en un texto con fuerza normativa; iv) con el principal objetivo de limitar el poder de los gobernantes.

Así, a manera de ejemplo, en la Carta Magna de 1215, Juan Sin Tierra fue obligado por sus barones a reconocerles ciertas prerrogativas —derecho a la propiedad y al patrimonio, a la libertad, al debido proceso—, como consecuencia de derrotas externas y la debilidad de su reinado.

Tras una revisión de la historia, "como ha ocurrido con la mayoría de los estatutos constitucionales, la Carta Magna no fue el fruto de la generosidad de un gobernante deseoso de reconocer o ampliar las libertades de sus súbditos" (Evangelista, 2015). Más bien, las constituciones se han caracterizado por ser *Cartas de Guerra*, en donde el vencedor impone condiciones, a través, por ejemplo, del reconocimiento de derechos, al vencido.

Pese a ello, los derechos humanos, como instrumento de control del pueblo frente a la Administración, pues, como se señaló, busca limitar "la arbitrariedad del poder político" (Evangelista, 2015), han tenido, a lo largo de la historia, mayor contenido discursivo, que jurídico. Superar la emotividad que suponen los derechos humanos, y acercarse a definiciones que les den forma es el primer paso para lograr su verdadera materialización.

Para eso, debemos proponernos ciertas preguntas, como ¿qué son los derechos humanos?, ¿cuáles son sus elementos esenciales?, ¿dónde están consagrados?, ¿cómo se protegen? Lo anterior, nos llevará a una definición formal de los derechos humanos.

Entender que los derechos humanos no son inherentes a la persona es la base. Abandonar el utópico mundo de la concepción iusnaturalista de los derechos humanos y aterrizar en definiciones que les den forma, es la mejor manera de protegerlos y buscar que se garanticen ese mínimo de derechos a la mayor cantidad de personas.

3.2.2. Es imposible proteger un derecho sin saber qué es y cuáles son sus elementos, esto es, sin delimitarlo. Más allá de la perspectiva iusnaturalista

La dogmática se ha limitado a definir los derechos humanos como aquellos que les corresponden a los seres humanos por el mero hecho de existir (Paine, 1791). Esa abstracta definición se ha nutrido a su vez de conceptos aún más etéreos, como la dignidad humana, la libertad, y la igualdad (Pérez Luño, 2003).

¿Cómo podemos proteger algo que no sabemos qué es?, ¿cómo podemos conceptualizar sobre la base de algo tan etéreo? Esa dificultad frente a la falta de consenso alrededor de la dignidad humana es estudiada a profundidad por Ernst Benda (2001). En

últimas, los derechos humanos buscan ser tanto, que terminan siendo poco, y muchas veces, nada.

Néstor Osuna (2017), señala que al fundar las definiciones que en líneas anteriores fueron analizadas como no formales, en conceptos no unívocos, y que, además, no provienen del derecho, como la dignidad, la libertad y la igualdad, estas, más que definiciones "son grandes ideas, lo cual no contribuye a acotar el objeto definido" (Osuna, 2017), además de que:

> […] minusvaloran la existencia de normas vigentes que establezcan los derechos a los que aluden, pues se satisfacen con que tales prerrogativas "merezcan" reconocimiento, o "deban ser" recogidas por los ordenamientos, con lo cual presentan falencias graves para un análisis propiamente jurídico, pues para el derecho no es equivalente, ni puede serlo, aquello que las leyes establecen a aquello que ignoran o deberían establecer. (pp. 346-347)

Es bajo ese entendido, que para Osuna (2017), los derechos humanos son "normas de rango supralegal, que le otorgan a alguien una situación de ventaja", por lo que "toda disposición constitucional o internacional que establezca para alguien una situación de ventaja es una disposición de derechos fundamentales o humanos"; además, y desde una perspectiva procesal, dicha situación de ventaja debe tener "una garantía judicial específica" (p. 347).

Los derechos humanos no son más que situaciones de ventaja que se otorgan a comunidades específicas en contextos históricos determinados. En últimas, los derechos humanos son privilegios. De ahí podemos entender que los derechos humanos no son inherentes a la persona —afirmación, además, empíricamente demostrable—, sino que son contextuales, y su materialización y efectividad dependen de su costo y voluntad política (Sunstein y Holmes, 2011).

3.2.3. ¿Para qué sirve consagrar derechos en tratados internacionales? Sistema Interamericano y Control de Convencionalidad

Los derechos humanos están consagrados en normas supralegales. Dentro del contexto actual, es entendida como norma supralegal aquella que esté consagrada en la Constitución o en los tratados

internacionales sobre derechos humanos. En efecto, estas prerrogativas son tan importantes para el hombre en sociedad, que deben estar positivizadas tanto nacional, como internacionalmente.

Por ello, desde la Declaración Universal de los Derechos Humanos de las Naciones Unidas, pasando por la Convención Americana sobre Derechos Humanos y aterrizando en la mayoría —por no decir que en todas— las Constituciones modernas, hay positivizados derechos humanos.

A nivel mundial fueron creados Sistemas de Protección de Derechos Humanos —universales y regionales— respondiendo a la necesidad de la protección de dichos derechos ante las posibles vulneraciones y arbitrariedades cometidas por los Estados directa, o indirectamente.

Dentro de los Sistemas Regionales de Protección de Derechos Humanos encontramos en Europa el Tribunal Europeo de Derechos Humanos, creado con el Convenio Europeo de Derechos Humanos de 1950; en África, el Sistema Africano de Derechos Humanos y de los Pueblos, creado con la Carta Africana de Derechos Humanos y de los Pueblos de 1981; y, en América, el Sistema Interamericano que se consolidó con la Convención Americana sobre Derechos Humanos, en adelante Convención ADH, que le dio vida a la Corte Interamericana de Derechos Humanos, en adelante Corte IDH.

El Sistema Interamericano de Derechos Humanos nace a partir de la Carta de la Organización de los Estados Americanos. Este está compuesto por la Comisión IDH y la Corte IDH, y se encargan de conocer los casos en que se aleguen violaciones a Derechos Humanos por parte de los Estados americanos.

Si bien hay 23 países, de los 35 que hay en nuestro continente, que ratificaron la Convención ADH (Argentina, Barbados, Bolivia, Brasil, Chile, Colombia, Costa Rica, Dominica, Ecuador, El Salvador, Granada, Guatemala, Haití, Honduras, Jamaica, México, Nicaragua, Panamá, Paraguay, Perú, República Dominicana, Surinam y Uruguay), hay 3 que no realizaron el acto de reconocimiento de competencia contenciosa de la Corte: Dominica, Granada y Jamaica. A estos países se suman aquellos que no realizaron la ratificación o la firma del tratado, como, por ejemplo, Estados Unidos.

Entendiendo entonces que la Corte IDH es el máximo intérprete de la Convención ADH, su actividad jurisdiccional en los casos contenciosos se vuelve vinculante para los Estados parte que hubiesen reconocido competencia contenciosa a la Corte. Es decir que, las sentencias de la Corte IDH y las interpretaciones a derechos en ella contenida, no solo son vinculantes para el Estado en litigio, sino para todos los Estados que hubiesen hecho el reconocimiento de competencia contenciosa a la Corte. Lo anterior, porque en sentido material, las sentencias se vuelven parte de la Convención misma, por lo que, al igual que los tratados, deben ser acatadas bajo el principio *pacta sunt servanda* (Yanguas Ramón, 2019).

¿Por qué es importante tener en cuenta ese ordenamiento convencional? Los ordenamientos internos tienen el deber de adaptarse a los estándares contenidos, tanto en los instrumentos válidos dentro del Sistema Interamericano, como en los tratados vinculantes y sentencias. Surge entonces, un deber en dos dimensiones; i) el deber de los Estados de observar que sus actuaciones se alineen a los parámetros interamericanos, y ii) el deber de los Estados de, en caso de existir colisión entre una norma dentro del ordenamiento interno y una norma convencional, aplicar aquella que favorezca en mayor medida a la persona, o aquella que esté mejor desarrollada, conforme al principio *pro homine*.

Esos deberes de los Estados se traducen en últimas en lo que se conoce como control de convencionalidad. Este control ha sido clasificado, tanto por la jurisprudencia, como por la doctrina, de acuerdo con quien lo realiza. El control de convencionalidad concentrado es ejercido directamente por la Corte IDH, en ejercicio de su facultad jurisdiccional, analizando las normas del ordenamiento interno del Estado y su coherencia con la Convención ADH.

El control de convencionalidad difuso está en cabeza del Estado, quien tiene el deber de verificar, de preferencia antes de cualquier actuación (por parte incluso, de la misma autoridad), o después (por la misma autoridad o por parte de autoridad judicial), que sus normas, bien sea legales o constitucionales, se ajusten a los estándares interamericanos.

3.2.4. El enfoque basado en derechos humanos

Comprendiendo el control de convencionalidad difuso, así como el principio del derecho internacional *pacta sunt servanda*, podemos señalar que los Estados tienen obligaciones internacionales concretas, que deben cumplir a través de sus órganos y entidades, específicamente frente al respeto y protección de derechos humanos. En principio, dicha obligación está en cabeza del Estado, como sujeto de derecho internacional. Es el Estado quien se obliga, por lo que radica en él, en principio, la responsabilidad de garantizar que sus actuaciones respeten derechos humanos, y de elaborar políticas públicas que, desde su formulación, tengan presente el enfoque basado en derechos humanos, en adelante EBDH.

Sin embargo, esa responsabilidad del Estado de garantizar que sus actuaciones respeten los derechos humanos se extiende a terceros, pues es su deber vigilar que, dentro de su territorio, se protejan y respeten los derechos humanos. Para eso, los Estados no solo deben desarrollar políticas con EBDH, sino también realizar planes y acciones tendientes a garantizar que terceros con operaciones dentro del territorio nacional actúen bajo la guía de los derechos humanos.

Entrando en materia, podemos concluir que se desprenden obligaciones, específicamente frente a la creación de tecnología, tanto por parte del Estado, como de empresas. Las Naciones Unidas, entendiendo lo anterior, elaboró una guía que incluye los principios rectores en las empresas y derechos humanos, que serán resumidos a continuación:

En primer término, el *deber general del Estado de proteger Derechos Humanos*. El Estado debe proteger a las personas contra violaciones de sus derechos humanos, de sí mismo y de terceros. Lo anterior, incluye las obligaciones de prevenir, investigar, sancionar y reparar violaciones de derechos humanos.

El Estado *tiene la obligación de advertir a terceros frente al deber que tienen de respetar Derechos Humanos*. El Estado tiene la expectativa de que las actuaciones y operaciones de terceros no vulnerarán derechos humanos, pero debe advertir que en caso en que estos sean violados, serán investigados, sancionados y tendrán el deber de reparar dichas violaciones.

El Estado tiene el *deber de proveer mecanismos judiciales efectivos*. El Estado debe garantizar mecanismos judiciales efectivos y otras soluciones en caso de violaciones a los derechos humanos.

Las compañías que diseñan IA, tienen el deber de respetar Derechos Humanos. Las compañías deber ser conscientes del alto impacto que pueden tener los softwares que desarrollen, por lo que deben respetar derechos humanos en sus diseños (ONU, 2011).

Lo anterior implica que los Estados tienen el deber de garantizar que las empresas del Estado, pero especialmente aquellas de naturaleza privada, diseñen programas impulsados por IA respetuosos de los derechos humanos.

Ello adquiere relevancia en el sentido de que son las empresas privadas, principalmente, quienes crean, desarrollan, utilizan e implementan sistemas de IA y, a su vez, es el Estado quien tiene el deber de garantizar que lo hagan de conformidad con los parámetros jurídicos, especialmente aquellos límites impuestos por los derechos humanos.

3.2.5. El diseño basado en derechos humanos

Debemos reconocer que la tecnología, específicamente la IA, jamás podrá ser objetiva. Los seres humanos, por naturaleza, no lo somos, ya que vemos el mundo conforme a las creencias que nos han inculcado, a la cultura que nos ha permeado en nuestro entorno social, y a todos los agentes externos a los que estamos expuestos en nuestras vidas. Nosotros somos los creadores de la IA; por lo tanto, esta replica nuestros sesgos. Reconocer lo anterior es el primer paso para trabajar en disminuirlos a su mínima expresión y propugnar que estos diseños respeten derechos humanos (*human rights centered design*, o *human rights by design*).

Tanto el Estado como las compañías que desarrollen tecnología, deben considerar a corto, mediano y largo plazo, el impacto de sus *softwares* en los derechos humanos, así como planear y hacer seguimiento de los efectos negativos, con el objetivo de mitigarlos. Se hace necesario, entonces, delimitar las etapas de creación de un sistema de inteligencia artificial, con el objetivo de generar estrategias específicas de protección de derechos humanos en cada una de ellas.

La creación de las IA comprende cuatro etapas importantes, i) diseño de las tecnologías y sistemas; ii) desarrollo, o fase de producción; iii) implementación, que es la distribución y uso de la IA y iv) evaluación e impacto, que incluye la medida y el análisis del impacto de tecnologías de IA (Guío Español, 2020).

Ahora bien, para garantizar un EBDH y poder analizar el impacto de la IA en los derechos humanos, es necesario estudiar las condiciones del entorno en donde se aplicará la IA.

La Universidad de Harvard publicó un artículo que propone un método de dos pasos para identificar el impacto positivo y negativo de introducir una IA a un entorno determinado (Raso *et al.*, 2018). Primero se debe considerar que puede haber implicaciones en derechos humanos, por lo que se debe evaluar la disponibilidad y efectividad de mecanismos institucionales encaminados a regular y reparar los impactos negativos en derechos humanos. Segundo, identificar el impacto de la IA; si esta mejora el funcionamiento de los derechos humanos, o, por el contrario, lo empeora.

En este segundo paso se debe verificar i) si los datos que se usan para entrenar la IA están sesgados, ii) el diseño del sistema y la injerencia de los sesgos y experiencias personales de los diseñadores, y iii) la interacción imprevista de la IA con su entorno, en la que posiblemente afecte los derechos humanos.

3.3. Estudio de caso: ROMA

La figura de la Fiscalía, como entidad encargada de investigar las conductas que son consideradas delitos, y en acusar en caso de encontrar mérito para ello ante los jueces y tribunales, se ha generalizado en la mayoría de los Estados modernos.

Aunque podría parecer utópico —o distópico, al estilo orwelliano—, la IA hasta cierto punto "puede reemplazar a los fiscales en el proceso de toma de decisiones" (Berbell, 2022). En China, la Fiscalía Popular de Shanghái Pudong implementó un sistema informático basado en IA capaz de elaborar escritos de acusación y presentar cargos contra sospechosos basándose únicamente en una descripción verbal, con una precisión, según reportan, superior al 97 % (Chen, 2021).

La nueva tecnología alivianaría la carga de los fiscales humanos y les permitiría centrarse solo en los casos más complejos. Además, el sistema de IA puede ejecutarse en una computadora de escritorio estándar (Chen, 2021).

El fiscal de inteligencia artificial fue entrenado usando 17 000 casos de la vida real desde 2015 hasta 2020 y es capaz de identificar y presentar cargos por los ocho delitos más comunes en Shanghái: fraude, fraude con tarjeta de crédito, robo, conducción peligrosa, lesiones intencionales, obstrucción de los deberes oficiales, dirigir una operación de juego y "buscar peleas", y "provocar problemas" (Chen, 2021).

En Argentina, la Fiscalía de la Ciudad Autónoma de Buenos Aires desarrolló Prometea, la primera IA de Latinoamérica al servicio de la Justicia (Corvalán, 2018). Prometea es un sistema de IA que prepara automáticamente dictámenes judiciales (Estevez *et al.*, 2020). El sistema busca "automatizar tareas reiterativas y la aplicación de IA para la elaboración automática de dictámenes jurídicos basándose en casos análogos para cuya solución ya existen precedentes judiciales reiterados" (p. 88).

Según el informe de avances, Prometea ha permitido a la Fiscalía incrementar la eficiencia de sus procesos de manera significativa: "la reducción de 90 minutos a 1 minuto (99 %) para la resolución de un pliego de contrataciones, de 167 días a 38 días (77 %) para procesos de requerimiento a juicio, de 190 días a 42 días (78 %) para amparos habitacionales con citación de terceros, entre otros" (Estevez *et al.*, 2020, pp. 10-11), lo que ha permitido que los empleados y funcionarios dedicados a dichas labores puedan enfocarse en casos más complejos (Estevez *et al.*, 2020).

En Chile, la Universidad de Chile desarrolló un sistema de IA capaz de construir una red de vínculos entre personas con historial delictivo, además de identificar a potenciales integrantes de bandas asociadas a hechos criminales, que será utilizado por la Fiscalía Nacional para detectar estructuras criminales (Universidad de Chile, 2023).

Colombia, a pesar de sus retrasos en materia tecnológica, no ha sido ajena a dicha realidad. Watson y PRiSMA, en la Fiscalía General;

Pretoria, en la Corte Constitucional, son algunos ejemplos de IA aplicada en la administración de justicia. Pero para llegar a ellos, es necesario señalar que lo que la sociedad colombiana está viviendo, en relación con el debate ético y jurídico de la implementación de IA en el sector público, no es algo nuevo.

En las décadas del ochenta, noventa y recién entrado el nuevo siglo, cuando se empezaron a introducir los primeros equipos de cómputo a la Rama Judicial, los jueces, casi al unísono, elevaron distintos temores: ¿seremos reemplazados por los computadores? ¿La información consignada es segura? ¿Realmente el juez está tomando la decisión si se utiliza una máquina?

Hoy, dos décadas después, los funcionarios ni siquiera logran concebir cómo todas las tareas que hoy se realizan con ayuda de las nuevas tecnologías, podían hacerse antes sin ellas: notificar personalmente, libros electrónicos, estados, las labores de sustanciación, las audiencias por medios virtuales, el envío de expedientes a otras dependencias.

Sin duda, la introducción de esas herramientas tecnológicas, entendidas como ello, como *herramientas*, revolucionaron e hicieron más eficiente la administración de justicia. La pregunta, para el lector, es: ¿no está pasando exactamente lo mismo con la IA? ¿No se estaría gestando, sin fundamento, un nuevo movimiento ludista?

En efecto, la introducción de herramientas potenciadas por IA ha generado la misma reticencia de los funcionarios de la Rama Judicial que se vivió en su momento cuando se empezaron a utilizar los equipos de cómputo. Surgieron las mismas dudas que han surgido a lo largo de la historia: ¿los humanos serán reemplazados por la tecnología, específicamente por la IA?

Si se responde esa pregunta de carácter predictivo con base en las vivencias de la humanidad, tendría que responderse que no. Si bien las máquinas de manufacturación en el siglo XIX empezaron a desempeñar funciones que antes cumplían, por ejemplo, los artesanos, esas máquinas requerían de personal calificado para operarlas. Si bien los computadores, desde 1980, facilitaron las tareas de los funcionarios de la Rama Judicial, son inútiles sin una persona que los utilice.

En esa misma línea, la IA sería inservible sin un humano que la cree, la programe y la utilice. Por ello, en el estado actual de los avances tecnológicos, sería impensable que la IA reemplace al funcionario judicial, generando la necesidad de centrar el debate en la utilización de la IA como herramienta al servicio de la justicia.

3.3.1. El furor de las RATs. Análisis comparado a partir de COMPAS y HART

Aunque la IA puede ser utilizada para una infinidad de tareas, en relación con el apoyo a la administración de justicia, una de las más controversiales y conocidas son las denominadas RATs (*risk assesment tools*). Son cada vez más utilizadas para asistir a los fiscales en las decisiones sobre la libertad de las personas, a través, como su nombre lo indica, de una evaluación de riesgo del procesado. Así, busca que el operador judicial cuente con un mayor volumen de información, lo que se traduce en elementos de juicio adicionales para mejorar la calidad de la decisión.

En Estados Unidos, por ejemplo, hoy se utilizan diferentes herramientas de valoración del riesgo de reincidencia en tres momentos procesales: i) *pre-trial release*, para decidir sobre medidas de aseguramiento o "libertad provisional"; ii) *sentencing*, para evaluar la necesidad de la pena privativa de la libertad y recomendar medidas alternativas si es del caso; iii) en la etapa de ejecución de penas, que incluye el tratamiento penitenciario, la supervisión y la libertad condicional.

En el primer escenario, las decisiones sobre el *pre-trial release*, existen más de 20 herramientas predictivas del riesgo de reincidencia, siendo la más común PSA (*Public Safety Asessment*), desarrollada por una entidad privada sin ánimo de lucro, Arnold Foundation (Notaro, 2023). En esa etapa procesal, algunos Estados de manera expresa "permiten el uso de algoritmos predictivos para las decisiones en tema de pre-trial release, otros lo recomiendan y algunos lo establecen como obligatorio" (p. 2023).

En el segundo momento procesal, la fase del *sentencing*, el juez tiene el deber de establecer el tratamiento sancionatorio a imponerse al condenado, para lo cual se utilizaban en 2015 más de 60 diferentes

RATs en los tribunales estadounidenses, dentro de las cuales la más conocida es COMPAS (Correctional Offender Management Profiling for Alternative Sanction) (Huq, 2019; Kehl *et al.*, 2017).

COMPAS produce una calificación de riesgo de las personas privadas de la libertad y su probabilidad de reincidencia criminal. El algoritmo COMPAS utiliza la cantidad de arrestos, el tipo de crimen cometido y las características sociodemográficas de los ofensores para formular un puntaje de riego de reincidencia (Minaggia, 2023).

Sin detenernos en COMPAS, pues no es el objetivo principal de este artículo, es necesario señalar que existen dos métodos utilizados para arrojar un resultado predictivo: los clínicos y los actuariales. Los métodos clínicos utilizan datos dinámicos, como la trayectoria académica de un individuo o su sometimiento a un tratamiento, por ejemplo, mientras que los actuariales usan datos estáticos como la edad. El método clínico se centra entonces en factores subjetivos e individuales, mientras que el método actuarial se enfoca en factores objetivos (Minaggia, 2023).

COMPAS, como salta a la vista, utiliza ambos métodos: clínicos, en el sentido de basarse en la cantidad de arrestos del individuo, y actuariales, pues estructura su predicción con base en las características sociodemográficas de los ofensores. El método actuarial, como se analizará a profundidad en párrafos posteriores, genera —entre muchos—, dos principales dificultades desde la perspectiva constitucional y convencional: i) potencializa los sesgos y; ii) des individualiza la restricción de la libertad que, en todo caso, debe ser personalísima.

En términos generales, la IA bajo el método actuarial podría determinar con base en criterios sospechosos de discriminación, como el sexo, la raza, el estrato socioeconómico, entre otros, lo que representó un mayor riesgo. Esto, a su vez, se traduciría en medidas severas, no por mis actuaciones, sino por las actuaciones previas de personas con características similares a las mías.

Sumado a lo anterior, el funcionamiento de COMPAS, desarrollado en 1998 por la empresa privada Equivant (antes Northpointe), se encuentra protegido por el secreto empresarial, lo que ha generado dificultades en su aplicación. El asunto fue explorado a profundidad

en el polémico caso State v. Loomis (Recent Cases, 2017), el Tribunal Supremo de Wisconsin abordó varios aspectos sobre la compatibilidad del uso de COMPAS en la fase de *sentencing* con los derechos del acusado.

Eric Loomis, condenado a seis años de prisión mediando la utilización de COMPAS por el juez, apeló el fallo invocando tres violaciones a su derecho al debido proceso. Aseguró, como ya se señaló, que la decisión sobre su tratamiento sancionatorio careció de valoración individual y que la inaccesibilidad al funcionamiento del algoritmo que fue utilizado para condenarlo, derivada del secreto empresarial que lo protege y hace imposible auditarlo, atentó contra su derecho de defensa, en relación con los principios de contradicción e igualdad de las partes. El secreto empresarial, en últimas, generó que, tanto la validez de la herramienta predictiva, como su resultado, fuesen incuestionables e inescrutables.

Finalmente, Loomis aseguró que el algoritmo, al considerar características como el sexo y la raza de la persona para emitir el puntaje había realizado una discriminación injusta.

El fallo del Tribunal de Wisconsin ha sido duramente criticado en la doctrina estadounidense (Beriain, 2018; Freeman, 2016; Lightbourne, 2017; Recent Cases, 2017; Israni, 2017) y extranjera (Martínez Garay, 2018) por los argumentos esgrimidos para desestimar las pretensiones del accionante.

Las RATs se han extendido a otras latitudes. En Reino Unido, por ejemplo, se utiliza el algoritmo HART (*Harm Assesment Risk Tool*) para finalidades similares, con la particularidad de que clasifica a los sospechosos en tres grupos, según su nivel de riesgo de reincidencia en los siguientes dos años, en elevado, moderado y bajo.

La herramienta ha sido utilizada por la policía de Durham para seleccionar a las personas con riesgo moderado de reincidencia, para que accedan a un programa de rehabilitación denominado "*Checkpoint*" (Notaro, 2023). Sin embargo, el algoritmo ha sido criticado por tener un sesgo en contra de la población que habita los barrios más pobres, pues entre las 34 variables que procesa para evaluar el riesgo se encuentra el código postal del acusado (Burgess, 2018; Oswald *et al*., 2018).

3.3.2. Ahora sí, ¿qué es ROMA?

Como se señaló, Colombia no ha sido ajena al desarrollo de este tipo de herramientas. La Fiscalía General de la Nación desarrolló PRiSMA en 2018: una IA creada con el propósito de reducir la reincidencia criminal, administrar mejor los escasos cupos carcelarios y hacer un uso más racional de las medidas de aseguramiento intramurales, intentando que estas se concentren en aquellos individuos con altos niveles objetivos de riesgo de reincidencia criminal.

La Fiscalía General de la Nación, tras un estudio empírico concluyó que, tanto para sustentar la necesidad, proporcionalidad y adecuación de la medida, como para la imposición de la misma por parte del juez, los juicios se basaban en predicciones a partir de la información incompleta y heterogénea disponible al momento de la audiencia, lo que podía conducir a errores (Fiscalía General de la Nación, 2018).

Dentro de esos errores, al momento de decidir sobre la detención preventiva de una persona, se podía: i) otorgar medida de aseguramiento a un imputado con bajo riesgo de reincidencia; ii) no otorgar medida de aseguramiento a un imputado con alto riesgo de reincidencia (Fiscalía General de la Nación, 2018).

Los resultados del estudio arrojaron que en muchas ocasiones (31.9 %) se dejaba en libertad a individuos con alto riesgo de reincidencia, o se dictaba medida de aseguramiento contra personas con bajo nivel de riesgo (56.1 %) (Fiscalía General de la Nación, 2018). Dichos errores podían ser mitigados en la medida en que, al momento de la audiencia de imposición de medida de aseguramiento, se contara con más información objetiva, homogénea y confiable (Fiscalía General de la Nación, 2018). Para ello, la Fiscalía ideó PRiSMA. El algoritmo utiliza un modelo de aprendizaje supervisado (*machine learning*), mediante el cual se predecía la probabilidad de reincidencia, dadas las características del ofensor, del último evento criminal y de los antecedentes criminales de cada individuo —método clínico—.

El estudio de la Fiscalía concluyó que la utilización de la herramienta PRISMA, como complemento de la información que tienen los fiscales en las audiencias de solicitud de imposición de medidas

de aseguramiento permitiría: i) manteniendo el mismo número de personas que son cobijadas con medida de aseguramiento intramural (aproximadamente 24.000 anuales), disminuir en un 25 % el número de eventos de reincidencia criminal en cualquier delito; ii) manteniendo el mismo nivel de reincidencia criminal, disminuir en 36 % el número de personas cobijadas con medida de aseguramiento intramural; iii) reducir notoriamente los errores en cuanto a, *a)* otorgar medida de aseguramiento a un imputado con bajo riesgo de reincidencia; *b)* no otorgar medida de aseguramiento a un imputado con alto riesgo de reincidencia.

El algoritmo PRiSMA se convirtió en el proyecto ROMA (Reporte Orientativo de Medidas de Aseguramiento). El proyecto ya no solo buscaba racionalizar las medidas de aseguramiento en virtud de la causal de riesgo de reincidencia, sino producir predicciones[1], sobre las otras dos causales de detención preventiva: riesgo de fuga y posible obstrucción a la justicia por parte del imputado.

A la fecha, el algoritmo de ROMA es funcional, está comprobado que aportaría sustancialmente a la política criminal del Estado y se está adelantando una evaluación de impacto de su utilización entre un grupo piloto de fiscales.

3.4. Analicemos ROMA

El rol que ha venido desempeñando la IA con respecto a la labor de la Fiscalía puede estudiarse desde dos aristas: la primera, a partir de la automatización de procesos, que permite evacuar de manera rápida y eficaz las labores reiterativas y de baja complejidad. La segunda, más compleja y problemática desde el punto de vista jurídico, a partir de la creación de algoritmos de predicción que facilitan la labor de investigación del crimen y la toma de decisiones en el proceso penal.

[1] Es una secuencia de filtros de grandes volúmenes de datos que arrojan un rango de relaciones de probabilidad aritmética.

En cuanto a lo primero, la IA supone un avance y una oportunidad de facilitar que la Fiscalía se centre en los asuntos realmente complejos, delegando en los sistemas de IA los temas de menor relevancia e impacto, a partir de la automatización de las labores repetitivas. La automatización de labores representa, lo que en su momento, representaron los computadores para la Rama Judicial.

Sin embargo, el presente análisis se centrará en la segunda arista, pues es la que mayores complejidades representa para el mundo jurídico: la creación de algoritmos predictivos (RATs).

Desde una perspectiva constitucional y convencional, hay tres elementos cruciales para tener en cuenta en este gran debate: i) la "individualización" o el carácter personalísimo de la responsabilidad penal; ii) la vulneración del derecho al debido proceso, específicamente en su componente de contradicción y defensa; iii) la vulneración del derecho a la igualdad y no discriminación.

3.4.1. La "individualización" o el carácter personalísimo de la responsabilidad penal.

En la doctrina penal se distingue entre el Derecho Penal de autor y el Derecho Penal de acto. En el derecho penal de autor, el sujeto responde por sus condiciones psicofísicas, por su personalidad, por su supuesta inclinación hacia el delito. El derecho penal de autor se funda en un criterio determinista, y el sujeto es castigado, no por haber cometido una conducta típica y antijurídica, sino por tener la potencialidad de cometerla. El autor, en síntesis, responde por su ser (Sánchez, 2011).

Por otro lado, en el Derecho Penal de acto, acogido por la mayoría de Estados democráticos por fundarse en la dignidad humana y en filosofías de corte liberal, el sujeto responde por sus actos conscientes y libres, es decir por la comisión de conductas conocidas y queridas por el mismo, previstas expresa y previamente en la ley como contrarias a bienes fundamentales de la sociedad y de sus miembros y que hacen a aquel merecedor de una sanción. En Colombia, a partir de distintos preceptos constitucionales, especialmente el artículo 29, y distintas normas convencionales, como el Artículo 8 de la Convención Americana Sobre Derechos Humanos, hemos

construido un Derecho Penal de autor alrededor del concepto de culpabilidad (Sentencia C-077, 2006).

En consecuencia, al momento de determinar el riesgo de que una persona pueda representar un peligro para la seguridad de la comunidad o de la víctima; o que pueda afectar el adecuado rumbo del proceso; o que exista un riesgo de fuga, es necesario que se tomen, única y exclusivamente, sus datos. Es decir, para que los RATs, en Colombia, sean constitucionales y convencionales, es necesario que se utilice exclusivamente el método clínico y, en ninguna circunstancia, se acuda al método actuarial.

El método actuarial, al basarse en correlaciones estadísticas a partir de sujetos con características similares, puede conducir a una "desindividualización" de la decisión en el resultado predictivo. Por ejemplo, la atribución a un sujeto de un nivel de riesgo elevado no provendría de un juicio realizado con referencia al individuo específico, sino que representaría una extensión a esta persona de datos estadísticos relativos a sujetos con características similares. En otras palabras, que otras personas con ciertas características demográficas —y que, en Colombia, son criterios sospechosos de discriminación— hayan delinquido en el pasado no implica que una persona que se les parece lo haga en el futuro.

3.4.2. La vulneración del derecho al debido proceso, específicamente en su componente de contradicción y defensa

El derecho al debido proceso, contenido principalmente en los artículos 29 de la Constitución Política y 8.º de la Convención Americana sobre Derechos Humanos, contiene, entre otras garantías, la inviolabilidad del derecho a la defensa (Velásquez, 2013).

El derecho a la defensa, en palabras de la Corte Constitucional, es la

> […] oportunidad reconocida a toda persona, en el ámbito de cualquier proceso o actuación judicial o administrativa, de ser oída, de hacer valer las propias razones y argumentos, de controvertir, contradecir y objetar las pruebas en contra y de solicitar la práctica y evaluación de las que se estiman favorables, así como ejercitar los recursos que la otorga. (Sentencia T-544, 2015, introducción)

En últimas, "el derecho de defensa garantiza la posibilidad de concurrir al proceso, hacerse parte en el mismo, defenderse, presentar alegatos y pruebas" (sección 4.1.4), y se concreta en dos derechos: el derecho de contradicción y el derecho a la defensa técnica.

Las RATs representan un nuevo reto relacionado con la falta de transparencia. Los creadores y desarrolladores, cuando pertenecen al sector privado, protegen sus creaciones o la propiedad sobre ellas a través del *secreto empresarial*.

El sistema predictivo impulsado por IA, al ser un complejo algoritmo y producto tecnológico, puede estar protegido por el secreto empresarial, como sucede con COMPAS. Como consecuencia de ello, los procesados, concretando el ejemplo a través de Eric Loomis, no pueden ejercer a plenitud su derecho de contradicción y defensa, al desconocer la forma en que funciona el sistema que está impactando la definición de su situación jurídica.

Por lo anterior, un algoritmo predictivo al servicio de la administración de justicia, impulsado por IA, solo podría ser constitucional y convencionalmente viable en Colombia si se garantiza la publicidad de los códigos fuente y toda la información relevante para la comprensión integral del funcionamiento del algoritmo. Esto puede ocurrir, bien sea porque se pacte de tal forma con el privado que desarrolle el algoritmo, o porque sea el mismo Estado quien lo haga.

3.4.3. La vulneración del derecho a la igualdad y no discriminación

En Colombia, la Corte Constitucional ha establecido que no ser discriminado, es un derecho protegido tanto constitucional como internacionalmente, derivado del derecho a la igualdad, consagrado en el artículo 13 de la Constitución (Sentencia T-572, 2017).

Para materializar el derecho a no ser discriminado, la Corte Constitucional ha desarrollado la teoría de los criterios sospechosos de discriminación e, incluso, ha establecido la presunción de discriminación. Frente a los primeros, ha considerado que, a partir del artículo 13 Superior, se desprende un mandato, consistente en la:

> […] prohibición de discriminación que implica que el Estado y los particulares no puedan aplicar un trato discriminatorio a partir de criterios sospechosos construidos a partir de -entre otras- razones de sexo, raza, origen étnico, identidad de género, religión u opinión política. (Sentencia T-572, 2017, sección 6.1)

En cuanto a la segunda, es un tema tan sensible para el Estado, que la discriminación pasa a presumirse, invirtiéndose la carga de la prueba. En consecuencia, "corresponde al presunto responsable de tales acciones desvirtuar la naturaleza discriminatoria de sus actos u omisiones" T-236 de 2023. Sentencia SU-440 de 2021. Sobre el particular, también consultar las sentencias T-077 de 2016, T-030 de 2017 y T-804 de 2014.

Aterrizando en la creación de modelos predictivos impulsados por IA, la mente humana, sesgada por naturaleza, juega un papel fundamental, por lo que es posible que el sistema predictivo de IA reproduzca no solo los sesgos de su creador, sino los factores de discriminación estructurales del entorno en donde será aplicada, poniendo en riesgo los derechos referenciados. Ello implicaría, no solo una reproducción de los sesgos, sino una potencialización de aquellos.

Por ello, los sesgos algorítmicos que pueden llegar a tener los sistemas de IA son una preocupación cada vez más grande. ¿Cómo pueden evitarse? Consideramos, en este espacio, empezar a hablar de la teoría del riesgo permitido de Claus Roxin (1997).

Hay que reconocer, entonces, que la actividad de creación algorítmica, en especial aquella de naturaleza predictiva, es esencialmente riesgosa. Para ello, es necesaria la consolidación de un marco normativo, de naturaleza regulatoria, que permita establecer los límites, a manera de *lex artis*, en materia de creación de algoritmos impulsados por inteligencia artificial.

En este punto, la regulación de IA no solo es necesaria por sí misma, sino para determinar el grado de responsabilidad penal de los creadores cuando se generen sesgos discriminatorios.

Para la construcción de ese marco regulatorio, es necesario un enfoque ético basado en derechos humanos, que permita que la IA,

desde su creación, pasando por su implementación y seguimiento, respete las garantías fundamentales de las personas (*human rights by design*) (Penney *et al*., 2017).

Ahora bien, en este punto, el de los sesgos, es relevante señalar que en este debate suele pasarse por alto que el ser humano, especialmente el juez o el fiscal, cargan a su vez con prejuicios ideológicos, políticos, sociales, económicos, emocionales y anímicos, que inciden en cada una de sus decisiones. Incluso, un estudio determinó que los jueces con hambre son más severos en sus decisiones (Danziger *et al*., 2011; Levav, 2011). Por ello, es necesario preguntarse si lo que se espera de la IA es una decisión objetiva e imparcialmente perfecta, o solo una mejor decisión que la que tomaría un humano evaluado bajo los mismos parámetros.

¿Existe un derecho a ser juzgado por un humano? La pregunta, en el contexto actual de los avances tecnológicos en Colombia, es todavía intrascendente. Aunque no existe tal derecho y no puede desprenderse de la garantía del juez natural, la IA, al menos en Colombia y en las circunstancias actuales, no busca reemplazar a los jueces, sino ser una herramienta para ellos, tal como lo advirtió la Corte Constitucional en un reciente fallo (Sentencia T-323, 2024). En él, reconoció las posibles bondades de la IA generativa, especializada en jurisprudencia y normatividad colombiana, pues garantizaría procesos más ágiles, al poner a la mano del servidor judicial información de manera fácil y especializada. Sin embargo, declaró que dichas herramientas deben utilizarse "como mecanismos que apoyan y que nunca pueden sustituir la labor humana en el servicio de justicia" (Sentencia T-323, 2024).

4. Conclusiones y recomendaciones

La IA ha llegado para quedarse en el mundo del derecho. Es el deber de la academia, del Estado y del sector privado garantizar su creación, implementación y seguimiento desde un enfoque basado en derechos humanos, a partir de la consolidación de un marco normativo que regule éticamente el debido desarrollo de algoritmos impulsados por IA.

Las tres grandes oportunidades que ofrece la IA en el derecho penal son la automatización de tareas, la predicción de comportamientos y la creación de nuevos contenidos gracias al *boom* reciente la inteligencia artificial generativa.

Específicamente, la IA y las RATs supone un gran desafío para el derecho. Están en riesgo derechos como la privacidad, el habeas data, la igualdad, la no discriminación, el debido proceso, especialmente en relación con el derecho de contradicción y defensa, la presunción de inocencia, el carácter personal la responsabilidad penal y la culpabilidad como elemento esencial del derecho penal de autor.

Sin embargo, con un marco normativo adecuado es posible construir algoritmos predictivos imparciales, libres de sesgos y auditables por los sujetos procesales para evitar decisiones injustas, discriminatorias o no explicables basadas en los datos procesados por estos sistemas.

Ese marco normativo, además de servir como base para la construcción de una IA respetuosa de los derechos humanos, puede también servir como parámetro para definir la *lex artis*, en relación con los riesgos permitidos, al momento de la creación y desarrollo de algoritmos predictivos, lo que será de utilidad para el derecho penal al momento de establecer responsabilidades en cabeza de los programadores.

Lo anterior, teniendo en cuenta que es imposible eliminar todos los sesgos de la IA, como es imposible erradicarlos de los humanos. La IA reproduce los sesgos de sus creadores, además de los factores de discriminación estructurales del entorno en donde será aplicada. Ello genera que sea una actividad naturalmente riesgosa. Sin embargo, si se estableciese un marco normativo claro, en relación con la creación y desarrollo de la IA, podrían determinarse aquellos riesgos aceptados, o permitidos, y así sancionar a los programadores que, irrespetando el marco normativo, generen riesgos no permitidos.

Para que un algoritmo predictivo de IA se ajuste a los parámetros normativos en Colombia, es necesario que, al menos, se garantice: (i) la publicidad de los códigos fuente y toda la información relevante para la comprensión del funcionamiento del algoritmo, para garantizar el derecho al debido proceso de las partes; (ii) la

utilización de métodos clínicos (y nunca actuariales) en donde se garantice la personalidad e individualización del análisis predictivo, para garantizar el derecho a la igualdad y no discriminación; (iii) la integridad del entrenamiento de la IA a partir de información fiable, que sea auditable por el público, materializando la transparencia y el derecho al debido proceso; (iv) un enfoque ético basado en derechos humanos en su creación, implementación y seguimiento (*human rights by design*) (Penney *et al*., 2017).

Todo ello depende, como se señaló, de la construcción de un marco normativo con un enfoque basado en derechos humanos, tanto interno como internacional sólido, que garantice el adecuado funcionamiento de los sistemas de IA.

Finalmente, frente a la postura de la Corte Constitucional, disentimos. En algún momento, como ha pasado en otras latitudes, la IA probará ser más eficiente, objetiva, menos costosa y, en general, más bondadosa, que los fiscales y los jueces. Es posible, entonces, que la IA sí pueda reemplazar a los humanos. Si puede reemplazarnos en el futuro, por supuesto que, hoy, puede hacernos la vida, tanto en la Fiscalía como en la Rama Judicial, mucho más fácil.

Dejando ello de presente, es claro que Colombia se encuentra lejos, muy lejos, de esa realidad, de ese temor, de que la IA reemplace humanos. El debate actual se encuentra en una etapa tan incipiente, que se ha centrado en si se puede, o no, utilizar ChatGPT, cuando los algoritmos impulsados por IA van mucho más allá, como se ha expuesto en el presente artículo.

Referencias

Amador Hidalgo, L. (1996). *Inteligencia artificial y sistemas expertos*. Universidad de Córdoba, Servicio de Publicaciones.

Ardila, R. (2011). Inteligencia. ¿Qué sabemos y qué nos falta por investigar? *Revista de la Academia Colombiana de Ciencias Exactas Físicas y Naturales*, *35*(134), 97-103. 10.18257/raccefyn.35(134).2011.2491

Benda, E. (2001). *Manual de derecho constitucional*. Marcial Pons.

Benítez, R., Escudero, G., Kanaan, S. y Cencerrado, A. (2014). *Inteligencia Artificial Avanzada.* UOC.

Berbell, C. (2022, 3 de enero). *Confilegal*. China inventa «un fiscal» de Inteligencia Artificial capaz de presentar cargos con un 97 % de precisión: https://bit.ly/3Wo5Wqh

British Broadcasting Corporation (BBC). (2018, 31 de diciembre). "AlphaZero nos supera de manera profunda": la computadora que genera su propio conocimiento y juega como un "superhumano". *BBC News*. https://bbc.in/3TcPPcW

Corte Constitucional de Colombia (2017, 13 de septiembre). Sentencia T-572, T-5.877.618 (Antonio José Lizarazo Ocampo, M. P.). https://tinyurl.com/mpfwn9j4

Corte Constitucional de Colombia (2024, 2 de agosto). Sentencia T-323, T-9.301.656 (Juan Carlos Cortés González, M. P.). https://tinyurl.com/mw9szyfh

Corte Constitucional de Colombia (2015, 21 de agosto). Sentencia T-544, T-4.895.508, (Mauricio González Cuervo, M. P.). https://tinyurl.com/nhjfy9pa

Corte Constitucional de Colombia. (2006, 8 de febrero). Sentencia C-077, D-5915 (Jaime Araujo Renteria, M. P.). https://tinyurl.com/44eptz6f

Corvalán, J. G. (2018). Inteligencia artificial: retos, desafíos y oportunidades – Prometea: la primera inteligencia artificial de Latinoamérica al servicio de la Justicia. *Revista De Investigações Constitucionais*, *5*(1), 295-316. https://doi.org/10.5380/rinc.v5i1.55334

Chen, S. (2021, 26 de diciembre). *South China Morning Post*. Chinese scientists develop AI 'prosecutor' that can press its own charges. https://bit.ly/49WdjIH

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters.

Estevez, E., Linares, S. y Fillottrani, P. (2020). *PROMETEA: Transformando la administración de justicia con herrramientas de inteligencia artificial*. Banco Interamericano de Desarrollo.

Evangelista, M. C. (2015). A 800 años de la Carta Magna inglesa de 1215. *Pensamiento Penal*, pp.1-20.

Ferrante, E. (2021). *Inteligencia artificial y sesgos algorítmicos. ¿Por qué deberían importarnos?* Nueva Sociedad, pp. 27-36.

Fiscalía General de la Nación. (2018). *Fiscalía General de la Nación*. Herramienta PRiSMA: perfil de riesgo de reincidencia para la solicitud de medidas de aseguramiento. https://bit.ly/49W2d6w

Flórez Ruiz, J. F. y Díaz Benito, C. O. (2021). *Videovigilancia con reconocimiento facial, inteligencia artificial y derechos humanos: ni apocalipsis ni utopía*. CETyS y LATAM Digital.

Forbes. (2023, 18 de julio). How does China's Approach to AI regulation differ from the US and EU? https://tinyurl.com/585j7ppw

Galton, F. (1869). *Hereditary genius*. Macmillan and co.

García Serrano, A. (2016). *Inteligencia artificial: Fundamentos, práctica y aplicaciones* . RC Libros.

Gates, B. (2023, 19 de diciembre). *Bill Gates: Por qué soy optimista sobre el futuro de la IA*. CNN: https://cnn.it/3yK3NMv

Guío Español, A. (2020). *Marco ético para la inteligencia artificial en Colombia.* Consejería Presidencial para asuntos económicos y transformación digital.

Harari, Y. N. (2023, 14 de septiembre). *Yuval Noah Harari and Mustafa Suleyman on the future of AI*, The Economist. https://econ.st/47gItuQ

Lara Gálvez, J. C. (2020). Perspectivas de derechos humanos en la implementación de marcos éticos para la inteligencia artificial. En C. Aguerre ed.), *Inteligencia Artificial en América Latina y el Caribe. Ética, gobernanza y políticas públicas* (pp. 34-66). CETyS.

Logically. (2019). *5 examples of biased artificial intelligence.* Logically.

Lowry, S. y Macpherson, G. (1988). A blot on the profession. *British medical journal (Clinical research ed.), 296*(6623), 657–658. https://doi.org/10.1136/bmj.296.6623.657

Martínez Garay, L. (2018). Peligrosidad, algoritmos y due process: el caso State v. Loomis. *Revista de Derecho Penal y Criminología, 3*(20), 485-502.

Minaggia, M. (2023). La inteligencia artificial en el derecho penal: la utilización de algoritmos predictivos. *Estudios sobre jurisprudencia* (pp. 278-298).

MIT Technology Review. (2019). *Facebook has been charged over housing ads that discriminate on race, color, and religion.* Technology review.

Morales, L., Agudelo, S., Montoya, M. y Montoya, A. (2021). Inteligencia artificial en el proceso pena: análisis a la luz del Fiscal Watson. *Pensamiento Jurídico,* (54), 147-164.

Notaro, L. (2023). Algoritmos predictivos' y justicia penal desde una perspectiva italiana y europea. En J. Peris y A. Massaro (eds.), *Derecho penal, inteligencia artificial y neurociencias* (pp. 191-214). Universita degli Studi Roma Tre.

ONU. (2011). *Guiding principles on business and human rights.* Office of the High Commissioner for Human Rights of the United Nations.

Osuna, N. (2017). Derechos humanos y libertades. En M. Correa Henao, *Lecciones de derecho constitucional, Tomo I* (pp. 343-384). Universidad Externado de Colombia.

Paine, T. (1791). *El derecho del hombre*Penney, J., McKune, S., Gill, L. y Deibert, R. (2017). Advancing human-rights-by-design in the dual-use technology industry. *Journal of International Affairs, 71*(2), 103-110.

Pérez Luño, A. (2003). *Derechos Humanos. Estado de derecho y constitución.* Tecnos.

Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C. y Kim, L. (2018). *Artificial intelligence & human rights: opportunities and risks.* Harvard Law School.

Recent Cases. (2017). Criminal Law - Sentencing guidelines - Wisconsin Supreme Court requieres warning before use of algorithmic risk assessments in sentencing - State v. Loomis, 881 N.W.2d 749 (Wis. 2016). *Harvard Law Review*, 1530-1537.

Romero, J., Dafonte, C., Gómez, A. y Penousal, F. (2007). *Inteligencia artificial y computación avanzada.* Santiago de Compostela.

Roxin, C. (1997). *Derecho Penal. Parte General.* Civitas.

Sánchez, E. (2011). *La dogmática de la teoría del delito. Evolución científica del sistema del delito.* Universidad Externado de Colombia.

Silberg, J. y Manyika, J. (2019). *Trackling Bias in Artificial Intelligence (and in humans).* Mckinsey Global Institute.

Sunstein, C. y Holmes, S. (2011). *El costo de los derechos. Por qué la libertad depende de los impuestos.* Altuna Impresores.

Universidad de Chile. (2023, 24 de marzo). Sistema basado en Inteligencia Artificial será aplicado por Fiscalía de Chile para detectar estructuras criminales, *Universidad de Chile.* https://bit.ly/4aUPbHK

Velásquez, F. (2013). *Manual de Derecho Penal. Parte General.* Ediciones Jurídicas Andrés Morales.

Yanguas Ramón, J. E. (2019). *Ius constitucionale commune* interamericano: desafío necesario para una protección más amplia de los derechos humanos. En E. Velandia (ed.), *Derecho procesal constitucional. Litigio ante la jurisdicción constitucional* (pp. 139-157). VC Editores.