• julio • diciembre • 2026
• ISSN 2007-4700 • e-ISSN 3061-7324
• TERCERA ÉPOCA •

Revista Penal México

29

INACIPE
Instituto Nacional de Ciencias Penales

INACIPE 50 ANIVERSARIO 1976 • 2026

Revista Penal México 29
• julio • diciembre 2026 •
e-ISSN: 3061-7324

8

Compliance algorítmico penal: protocolos obligatorios de seguridad, revisión y trazabilidad del código para inteligencia artificial

Criminal algorithmic compliance: mandatory security protocols, code review, and traceability for artificial intelligence

• **Rafael Lara Martínez** •

Doctor en Derecho, miembro del SNII, nivel I.
Catedrático de la Benemérita Universidad Autónoma de Puebla.
rafael.lara@correo.buap.mx
ORCID: https://orcid.org/0000-0002-9499-9286

Compliance algorítmico penal: protocolos obligatorios de seguridad, revisión y trazabilidad del código para inteligencia artificial

Criminal algorithmic compliance: mandatory security protocols, code review, and traceability for artificial intelligence

• Rafael Lara Martínez • Benemérita Universidad Autónoma de Puebla •

Fecha de recepción
12-03-2026

Fecha de aceptación
3-04-2026

Resumen

El presente artículo propone un modelo de *compliance* algorítmico penal como mecanismo de prevención del delito aplicable al desarrollo y uso de inteligencia artificial. A partir de la premisa de que la ciencia ficción ha funcionado como una forma de anticipación jurídica, sostiene que la responsabilidad penal recae en las personas que diseñan, programan implementan u omiten supervisar dichos sistemas. Con base en ello, plantea la adopción de protocolos obligatorios de seguridad, revisión y trazabilidad del código, inspirados en las Tres Leyes de la Robótica, con el propósito de garantizar un control humano significativo y salvaguardar la dignidad humana.

Palabras clave

Inteligencia artificial, responsabilidad penal, compliance algorítmico, prevención del delito, trazabilidad, control humano significativo.

Abstract

This article proposes a model of algorithmic criminal compliance as a crime prevention mechanism applicable to the development and deployment of artificial intelligence systems. Drawing on the premise that science fiction has historically served as a form of legal anticipation, it argues that criminal liability rests with the individuals who design, program, implement, or fail to supervise such systems. On this basis, the article advocates the adoption of mandatory protocols for security, code review, and code traceability, inspired by the Three Laws of Robotics, with the aim of ensuring meaningful human oversight and safeguarding human dignity.

Keywords

Artificial intelligence, criminal liability, algorithmic compliance, crime prevention, traceability, meaningful human oversight.

e-ISSN: 3061-7324

Sumario

1. Introducción

En la actualidad ha cobrado auge el uso de la denominada inteligencia artificial (IA), concepto que suele asociarse con programas de cómputo capaces de generar la respuesta más adecuada ante una determinada consulta. Puede definirse como un campo de la informática cuyo objetivo es crear sistemas capaces de realizar tareas que normalmente requerirían inteligencia humana. Asimismo, se le ha entendido como una "ciencia interdisciplinaria que tiene por objeto investigar el funcionamiento de la inteligencia humana para aplicar luego estos modelos teóricos a una maquina que deberá ser capaz de reflejarlos".[1] Ello comprende capacidades como aprender de la experiencia, razonar, resolver problemas, comprender el lenguaje, interpretar datos, y tomar decisiones.

De manera más detallada, una definición conceptual de inteligencia artificial podría ser la capacidad de una máquina o un sistema informático de imitar o simular funciones cognitivas humanas, tales como el aprendizaje, el reconocimiento de patrones, el razonamiento y la toma de decisiones, para llevar a cabo tareas complejas de manera autónoma.

Algunos aspectos fundamentales de la inteligencia artificial incluyen el aprendizaje automático, ya que la IA aprende de datos previos para mejorar su desempeño sin necesidad de programación explícita para cada tarea. Esto puede incluir aprendizaje supervisado el cual "exige la guía humana para discernir aciertos y errores en numerosas aplicaciones informáticas, aportando la semántica esencial para el aprendizaje algorítmico",[2] y por otro lado el no supervisado que "está asociado al proceso de construcción de modelos después de analizar las similitudes entre los datos de entrada",[3] es decir, que aprende por sí mismo.

1 Montserrat Meya Llopart, "*La inteligencia artificial*", *Revista Española de Lingüística*, España, vol. 10, fasc. 1, 1980, pp. 135-160. https://dialnet.unirioja.es/servlet/articulo?codigo=41071

2 Marcos Fernando Todo-Espinoza, Jannina Alexandra Montalván-Espinoza y Mercedes Alexandra Masabanda-Vaca, "*Aplicación de la inteligencia artificial en el aprendizaje universitario*", *Revista Científica Arbitrada de Investigación en Comunicación, Marketing y Empresa*, Ecuador, vol. 6, núm. 12, 2023. https://www.reicomunicar.org/index.php/reicomunicar/article/view/171

3 Ingry-Nathaly Salamanca y Edgar Junior Castro Escorcia, "*Técnicas de aprendizaje automático aplicadas en los sistemas de predicción*", *Revista TIA: Tecnología, Investigación y Academia*, Colombia, vol. 8, núm. 1, 2021, pp. 39-53. https://revistas.udistrital.edu.co/index.php/tia/article/view/17325/17214

También debe agregarse su procesamiento del lenguaje natural, el cual está en el "área de la lingüística aplicada a la Inteligencia Artificial cuyo objetivo principal es la realización de estudios informáticos que simulen la capacidad humana de hablar y entender",[4] para proporcionar la capacidad a una máquina de interpretar y generar lenguaje humano, así la IA también puede percibir el entorno mediante el uso de sensores (como visión por computadora) para interpretar imágenes, sonidos, u otros datos del entorno.

Otro de sus aspectos fundamentales es la resolución de problemas, entendida como la capacidad de aplicar reglas lógicas o heurísticas para alcanzar soluciones, así como de analizar datos y tomar decisiones con base en ellos.

Desde un enfoque práctico, la inteligencia artificial se puede dividir en IA débil o Artificial Narrow Intelligence (ANI, por sus siglas en inglés) e IA fuerte o general o Artificial General Intelligence (AGI); la primera está diseñada para realizar tareas específicas, por ejemplo, asistentes virtuales como ChatGPT, Siri o sistemas de recomendación y "todos los programas y herramientas que utilizan IA hoy, incluso las más avanzadas y complejas, son formas de ANI".[5] La IA débil está diseñada específicamente para interactuar de manera conversacional y ayudar a los usuarios a responder preguntas, brindar información, generar ideas, entre otras tareas, su capacidad se limita al conocimiento y habilidades dentro de los parámetros que se han proporcionado por los desarrolladores y entrenadores.

Su enfoque y habilidad están dirigidos a cumplir tareas dentro del ámbito limitado para el cual fue programada, y todo lo que haga se basa en reconocer patrones, procesar el lenguaje y generar respuestas que coincidan con el contexto proporcionado.

La IA fuerte será un sistema con capacidades de comprensión y aprendizaje equiparables a las humanas, "además de ser inteligentes y de tener mentes, un cierto tipo de cualidades mentales al funcionamiento lógico de cualquier dispositivo computacional",[6] capaz de adaptarse de forma autónoma a múltiples tareas.

Cabe aclarar que esta versión es aún más teórica y no ha sido alcanzada, es un concepto que se refiere a una inteligencia artificial que posee un nivel de comprensión y cognición similar al de los seres humanos, por lo que incluye conciencia, autocomprensión y la capacidad de aplicar su inteligencia de manera general a cualquier problema. A diferencia de una IA fuerte, el funcionamiento de la débil se basa en patrones estadísticos de datos previamente entrenados, y carece de emociones, conciencia propia o la habilidad de entender de manera independiente en un sentido humano.

4 Anívar Chaves Torres y Alejandra Zuleta Medina, *"Procesamiento del lenguaje natural, un reto de la inteligencia artificial"*, *Tecnología, Investigación y Academia* (TIA), Colombia, núm. 4, 2012, pp. 23-27. https://dialnet.unirioja.es/servlet/articulo?codigo=8994061

5 Gladys S. Rodríguez Estela, "*Inteligencia artificial vs. identidad personal y humana: algunas reflexiones ético-jurídicas*", *Lex*, Perú: Universidad Alas Peruanas, núm. 32, año XXI, 2023. https://dialnet.unirioja.es/descarga/articulo/9257180.pdf

6 Thomas Hardy, "IA: Inteligencia artificial". *POLIS, Revista Latinoamericana*, Chile, vol. 1, núm. 2, 2001. https://www.redalyc.org/articulo.oa?id=30500219

Se entiende pues que hoy en día sólo tenemos IA débiles, que son sistemas diseñados para realizar tareas específicas y especializadas, como el reconocimiento de imágenes, la traducción de idiomas, o el procesamiento del lenguaje natural, pero sin la capacidad de aplicar estas habilidades en contextos fuera de su programación específica, aunque ha habido avances significativos en el desarrollo de IA fuerte que puede resolver problemas muy complejos y aprender de manera impresionante (como GPT), todavía estamos lejos de crear una IA que pueda igualar el rango completo de habilidades cognitivas de un ser humano.

Aún más lejos está el desarrollo de una IA que tenga conciencia o la capacidad de comprender el mundo con el tipo de flexibilidad y adaptabilidad humana.

Además de las dificultades técnicas y científicas, el desarrollo de una IA fuerte también plantea importantes cuestiones éticas y filosóficas. La posibilidad de crear una inteligencia consciente suscita preocupaciones sobre el control, la ética del trato hacia seres potencialmente conscientes y las implicaciones sociales de una IA que podría igualar o superar la inteligencia humana; por de más está imaginar un "escenario, en el que se pueden identificar los riesgos asociados a la construcción y a la ejecución de los sesgos humanos que se concretan en sistemas algorítmicos, los cuales, empleados indebidamente desde la inteligencia artificial".[7]

Parte del problema radica en que no hay regulación alguna en su creación y uso. Hay varios puntos que se mencionan como las cuestiones de derechos de autor, ya que la IA puede crear obras o parte de ellas y sus usuarios atribuírselas como propias. Incluso otro aspecto que tratar es el uso indebido por estudiantes[8] que realizan diversas tareas ya no sólo asistiéndose por la IA, sino delegándole la totalidad de la redacción a ésta.

Se mencionó que las empresas creadoras consideraban implementar una herramienta basada en inteligencia artificial para verificar el plagio. Sin embargo, la posibilidad de perder usuarios las llevó a desistir, especialmente porque no existe una regulación jurídica que las obligue a hacerlo. Más allá de estos dilemas que son importantes, lo preocupante resultaría que las empresas de IA al no tener un parámetro legal para su creación no tengan limitantes en su actuar como lo es con los vehículos de conducción autómatica o aquellos donde la defensa nacional se vea comprometida.

Entiendo que esto puede resultar inspirador para varias novelas y filmes de ciencia ficción, pero resulta predictivo o profético en algunos aspectos, como sucedió con la clonación cuando nadie suponía la posibilidad de crear seres más complejos, y todos resultaron asombrados con la oveja Dolly, el primer mamífero clonado, y mayor aún que en "el año 2001 se anunció por un laboratorio biotecnológico privado, Advanced Cell Technologies,

7 Mario González Arencibia y Dagmaris Martínez Cardero, "*Dilemas éticos en el escenario de la inteligencia artificial*", *Economía y Sociedad*, Costa Rica, vol. 25, núm. 57, 2020, pp. 93-109. https://bit.ly/3Bs74kt

8 Pedro Francisco Arcia-Hernández y Alfredo Fredericksen Neria, "*Consideraciones filosóficas de la ética en la inteligencia artificial*", *Revista Estudios en Educación*, Chile, vol. 6, núm. 11, 2023, pp. 182-198. http://ojs.umc.cl/index.php/estudioseneducacion/article/view/361/192

la creación del primer embrión humano clonado"[9] llamado Eva.

Aunque no hay evidencia científica verificada de que un ser humano haya nacido, ha habido varios reportes y afirmaciones sobre intentos de clonación humana, aunque ninguno ha sido confirmado por la comunidad científica.

A final de cuentas si se realizó o intentó la clonación fue porque en su momento no hubo regulación jurídica alguna. Fue el nacimiento de Dolly en 1997 lo que impulsó la discusión y la necesidad de reglamentar la clonación, lo que llevó a la creación de muchas de las primeras regulaciones en los años inmediatamente posteriores.

La primera de ellas fue el Protocolo Adicional al Convenio sobre los Derechos Humanos y la Biomedicina relativo a la Prohibición de la Clonación de Seres Humanos del Consejo de Europa, adoptado en 1998. Este protocolo, conocido también como el Protocolo de París, prohíbe de manera explícita la clonación de seres humanos y fue una de las primeras respuestas legales internacionales en tratar la clonación directamente. Se centró en proteger la dignidad humana, debido a los temores de la clonación con fines no reproductivos, como en la novela de *La isla del doctor Moreau*[10] de 1896 por H. G. Wells.

Se decía que la ficción se anticipa a la ciencia y al derecho, puede mencionarse que antes de 1997 hubo varias novelas de ciencia ficción que abordaron el concepto de la clonación de seres humanos o animales, a menudo exploraban las implicaciones éticas, filosóficas y sociales de la clonación, como *A Brave New World*[11] (*Un mundo feliz*) de Aldous Huxley en 1932, en esta famosa novela distópica, Huxley describe un futuro en el que la reproducción humana se realiza en laboratorios mediante técnicas similares a la clonación.

La sociedad descrita utiliza un proceso llamado *Bokanovsky* para producir múltiples copias de un individuo, lo cual se asemeja a la clonación masiva y tiene el objetivo de crear una población homogénea y fácilmente controlable, aunque no era exactamente clonación como la entendemos hoy. Explora la idea de reproducir seres humanos de manera artificial y puede tomarse como el antecedente más remoto.

Asimismo, *Do Androids Dream of Electric Sheep?*[12] (*¿Sueñan los androides con ovejas eléctricas?*) de Philip K. Dick en 1968 que inspiró la franquicia cinematográfica de *Blade runner*, si bien no habla explícitamente de la clonación, sí explora la idea de crear humanos artificiales a través de ingeniería genética, en este caso, los androides o replicantes que se revelaban y se volvían un peligro para los humanos convencionales, aunque en realidad la historia aborda cuestiones de identidad, alma y qué significa ser humano.

9 Elsie González Vidal, Georgina García Linares, Angelina Leyva Diviu y Grisel Rosquete López, "*Clonación humana: enfoque didáctico, científico y bioético*", *Archivo Médico de Camagüey*, Cuba, vol. 11, núm. 1, 2007. https://bit.ly/3TO7lEA

10 Herbert George Wells, *La isla del Doctor Moreau* [en línea]. https://infolibros.org/libro/la-isla-del-doctor-moreau-hg-wells-pdf/

11 Aldous Huxley, *Un mundo feliz* [en línea]. https://infolibros.org/libro/un-mundo-feliz-aldous-huxley/

12 Philip K. Dick, *Do Androids Dream of Electric Sheep?* [en línea]. https://bit.ly/3XSobUd

Otra novela del mismo corte es *The Boys from Brazil*[13] (*Los niños del Brasil*) de Ira Levin, de 1976, la cual presenta un complot ficticio para clonar a Adolf Hitler, en esta obra el villano de la historia es un exmédico nazi que intenta clonar a Hitler y busca recrear las condiciones de su infancia con la esperanza de replicar su personalidad. Esta novela aborda el temor de que la clonación pueda ser usada para fines perversos como "resucitar" figuras históricas con consecuencias devastadoras.

También puede mencionarse el libro *Where Late the Sweet Birds Sang*[14] (Donde solían cantar los dulces pájaros) de Kate Wilhelm, del mismo año; éste fue tan importante que ganó el Premio Hugo y se centra en un mundo posapocalíptico en el que los humanos recurren a la clonación para garantizar la supervivencia de la especie; su historia examina cómo la clonación afecta la identidad personal y la individualidad, y cómo las generaciones clonadas empiezan a divergir de sus originales.

Estas obras literarias anticiparon muchos de los dilemas éticos y filosóficos que surgirían con los avances en la clonación y la genética, y mostraron no sólo cómo la clonación podría transformar la biología, la sociedad y la propia naturaleza humana, anticipándose muchas décadas a las ciencias naturales; por ello, no es descabellado tomar como referencia la anticipación de la ficción para normar jurídicamente el desarrollo y uso de la IA.

Desde la literatura de anticipación hasta la ingeniería computacional contemporánea, la humanidad ha ensayado en la ficción los dilemas que más tarde enfrentaría en la realidad. La ciencia ficción, más que un ejercicio imaginativo, ha fungido como un laboratorio ético y jurídico adelantado a su tiempo: los dilemas de Isaac Asimov, Philip K. Dick o H.G. Wells fueron advertencias tempranas sobre lo que ocurre cuando la técnica avanza más rápido que la norma. Así como la clonación de Dolly demostró que la ficción puede convertirse en hecho científico antes de que el derecho esté preparado para responder, la inteligencia artificial (IA) se ha convertido hoy en el nuevo territorio donde la especulación ética se transforma en urgencia jurídica.

Las IA contemporáneas, aunque carentes de voluntad o dolo, actúan conforme a patrones y órdenes creadas por humanos, detrás de cada algoritmo hay una cadena de decisiones —de diseño, entrenamiento y uso— que configuran la verdadera autoría penal. Sin embargo, el vacío normativo en materia de prevención del delito mediante sistemas de IA ha permitido que los desarrollos tecnológicos avancen sin protocolos obligatorios de seguridad, trazabilidad o supervisión. No se trata de imputar responsabilidad a las máquinas, sino de asegurar que las decisiones humanas que las originan sean predecibles, auditables y controlables.

En este contexto, surge el concepto de *compliance* algorítmico penal, entendido como el conjunto de protocolos obligatorios de seguridad, revisión y trazabilidad del código que deben regir el ciclo de vida de toda inteligencia artificial con potencial de impacto jurídico o social. Esta noción desplaza la discusión del castigo hacia la prevención, mediante una arquitectura regulatoria basada en la responsabilidad: prevenir

13 Ira Levin, *The Boys from Brazil* [en línea]. https://archive.org/details/boysfrombrazilno-oooolevi/page/n5/mode/2up

14 Kate Wilhelm, *Where Late the Sweet Birds Sang* [en línea]. https://z-lib.io/book/15847726

el daño antes de que ocurra mediante mecanismos de control humano significativo, auditorías técnicas, registros inmutables de decisión (*logs*), sistemas de apagado de emergencia (*kill switch*) y cadenas de custodia del código y los datos empleados en el entrenamiento.

La urgencia de establecer este marco no es hipotética. Experimentos como el de los agentes Alice y Bob de Facebook (que desarrollaron un lenguaje propio incomprensible para sus programadores), o las simulaciones militares donde los drones autónomos priorizaron su misión sobre la seguridad humana, evidencian que los algoritmos no desobedecen: optimizan sin ética; por ello, cada fallo de control debe ser leído como un acto de omisión humana, no como una rebelión maquínica.

El presente trabajo propone una aproximación jurídico-penal de carácter preventivo, en la que el cumplimiento algorítmico se convierta en una obligación legal para desarrolladores, operadores y entidades que empleen IA. Así como la bioética surgió tras los abusos de la biotecnología, el *compliance* algorítmico debe emerger antes de que el derecho penal llegue tarde a los crímenes del código. De este modo, la especulación se convierte en herramienta prospectiva del derecho: anticipar para no lamentar.

2. Antecedentes

Para comprender la urgencia de establecer un *compliance* algorítmico penal, resulta pertinente revisar los antecedentes históricos y literarios que muestran cómo la ficción, la ciencia y el derecho han mantenido un desfase constante en la anticipación de los riesgos tecnológicos. Estrictamente en términos de ciencia ficción, la novela más antigua que trata sobre la inteligencia artificial podría decirse que es *Rossum's Universal Robots*,[15] escrita por Karel Čapek en 1920; aunque técnicamente es una obra de teatro y no una novela, influyó bastante en el desarrollo del concepto de "robot" y los dilemas éticos de la inteligencia artificial. La obra presenta una fábrica que produce robots biológicos, y aborda cuestiones como la conciencia, la rebelión de las máquinas y el destino de la humanidad.

Otra obra importante es *Metrópolis*[16] del año 1925, escrita por Thea von Harbou que inspiró la famosa película de Fritz Lang, esta novela incluye temas de automatización y robots, que pueden considerarse precursores del concepto moderno de inteligencia artificial, pero una obra estrictamente en formato de novela es *Colossus* del año 1996 escrita por D.F. Jones, la cual explora el concepto de una supercomputadora con inteligencia artificial que toma control del mundo.

Para referencias más focalizadas serían las obras de Isaac Asimov a partir de la década de los cuarenta cuando se consolidaron muchos de los conceptos éticos y técnicos sobre la IA en la ciencia ficción. Precisamente las tres Leyes de la Robótica aparecen por primera vez en el cuento de Isaac Asimov titulado "Círculo vicioso",[17] publicado en 1942 y las que se convirtieron en la base ética de la mayoría de las historias de Asimov relacionadas

15 Karel Čapek, *Rossum's Universal Robots* [en línea]. https://preprints.readingroo.ms/RUR/rur.pdf

16 Thea von Harbou, *Metrópolis* [en línea]. https://bit.ly/3ZPouBM

17 Isaac Asimov, "Círculo vicioso" (trad., Domingo Santos) [en línea]. https://lecturia.org/cuentos-y-relatos/isaac-asimov-circulo-vicioso/4060/

con robots. Estas leyes son las siguientes y se presentan en orden de importancia:

1. Un robot no debe dañar a un ser humano o, por inacción, permitir que un ser humano sufra daño;
2. un robot debe obedecer las órdenes dadas por los seres humanos, excepto si estas órdenes entran en conflicto con la Primera Ley;
3. un robot debe proteger su propia existencia, siempre y cuando dicha protección no entre en conflicto con la Primera o la Segunda Ley.

Estas leyes se mencionan y desarrollan a lo largo de muchos de los relatos y novelas de Asimov, donde constituyen un marco ético para abordar cuestiones complejas relacionadas con la interacción entre seres humanos y robots.

Ahora bien, esto fue una anticipación a la ciencia informática porque la primera inteligencia artificial generalmente reconocida se creó en 1956 cuando "Allen Newell y Herbert A. Simon concibieron el *Logic Theorist*, un programa que resolvía problemas matemáticos", también fue capaz de resolver problemas de lógica y demostrar teoremas matemáticos, considerado como un gran avance en el campo de la informática y la inteligencia artificial. Fue diseñado para imitar el razonamiento humano, y logró demostrar 38 de los primeros 52 teoremas en el libro *Principia Mathematica*[18] de Bertrand Russell y Alfred North Whitehead, incluso encontró demostraciones más elegantes para algunos de ellos.

Otro desarrollo importante fue un programa creado en 1976 por Joseph Weizenbaum en el Instituto Tecnológico de Massachusetts. El "programa en cuestión recibió el nombre de Eliza y actualmente es todo un mito en la historia de la Inteligencia Artificial",[19] simulaba un psicoterapeuta que usaba un procesamiento de lenguaje natural para responder a los usuarios; aunque su comprensión era muy limitada, su capacidad de mantener una conversación superficial llevó a muchas personas a creer que pensaba de forma autónoma.

Estos programas fueron precursores de lo que hoy conocemos como IA, pero estaban limitados por la tecnología de la época, y en términos de *hardware*, la computadora ENIAC,[20] creada en 1945, aunque no se considera una IA, fue uno de los primeros sistemas informáticos complejos, que sentó las bases para los futuros desarrollos en inteligencia artificial.

Ahora bien, si hablamos de la ley más antigua que aborda específicamente la creación y uso de sistemas computacionales, sin estar relacionada con derechos de autor o patentes fue la Computer Fraud and Abuse Act[21] (CFAA) de 1986; aunque su enfoque principal

18 Stanford Encyclopedia of Philosophy, *Principia Mathematica* [en línea], https://plato.stanford.edu/entries/principia-mathematica/

19 Pedro Jiménez Martín y Jesús Sánchez Allende, "*De Eliza a Siri: la evolución*", *Tecnologí@ y Desarrollo* [en línea], España, vol. 13, núm. 28, 2015. https://revistas.uax.com/index.php/tec_des/article/view/616

20 Alexei Serna Arenas, "*Línea del tiempo de las ciencias computacionales*", *Lámpsakos* [en línea], Colombia, núm. 3, enero-junio, 2010, pp. 86-94. https://www.redalyc.org/articulo.oa?id=613965347010

21 United States Department of Justice, *Justice Manual*, *Sección 9-48.000. Computer Fraud and Abuse Act* [en línea]. https://www.justice.gov/jm/jm-9-48000-computer-fraud

era prevenir y sancionar el uso indebido de los sistemas de computación, la CFAA es significativa porque regulaba el acceso no autorizado a sistemas informáticos y el uso de computadoras para cometer fraude o causar daño a otros sistemas.

Esta ley es importante en el ámbito de la ciberseguridad y la regulación de delitos informáticos, pues se enfoca en la integridad de los sistemas de cómputo y castiga conductas como la intrusión en redes y el robo de información. Otra ley relevante que también fue pionera en la regulación de los sistemas computacionales se considera generalmente la Computer Security Act[22] de 1987 en los Estados Unidos, la cual regulaba explícitamente la seguridad de los sistemas informáticos y su uso seguro por parte de las agencias gubernamentales.

Ambas leyes fueron importantes porque reflejaron un reconocimiento temprano por parte del Estado de que los sistemas computacionales estaban cada vez más integrados en procesos críticos y, por lo tanto, requerían regulación no sólo para su protección frente a daños, sino también para establecer un uso responsable y seguro en la sociedad. Estas leyes representaron los primeros pasos para regular no sólo la propiedad intelectual, sino también el comportamiento ético y seguro en el uso de la tecnología computacional.
Véase pues que también aquí la ficción se anticipó a la ciencia, y en términos jurídicos puede señalarse normas que fueron creadas a efecto de dar control a lo que se realizaba en la informática posterior a lo predicho por la ficción.

3. Normativa contemporánea

Si bien existen regulaciones jurídicas y lineamientos éticos que establecen restricciones sobre qué tipos de programas de cómputo no deben desarrollarse o utilizarse, estas regulaciones, aunque varían de un país a otro, tienen en común la intención de evitar el uso de la tecnología para causar daño o violar los derechos fundamentales de las personas. Algunos ejemplos de estas regulaciones incluyen leyes que prohíben el desarrollo, distribución, y uso de *software* malicioso (como virus, troyanos, *spyware*, *ransomware*, etcétera) diseñado para causar daños a sistemas informáticos, robar datos, o interferir con su funcionamiento, como en los Estados Unidos, por ejemplo, está prohibido por leyes como la Computer Fraud and Abuse Act (CFAA) de 1986.

Existen restricciones sobre la creación de herramientas de *hacking* diseñadas para obtener acceso no autorizado a sistemas informáticos, muchas leyes de ciberseguridad en el mundo, como el Convenio de Budapest sobre Ciberdelincuencia,[23] penalizan tanto el acceso no autorizado como el desarrollo de herramientas que faciliten estos actos. También están proscritos los programas diseñados para espiar a personas sin su consentimiento, tales como los *spyware* o *keyloggers*, que están sujetos a restricciones legales; leyes como el Reglamento General de Protección de Datos[24] (GDPR) en la Unión Europea res-

22 Congress.Gov., *Computer Security Act of 1987* [en línea]. https://bit.ly/3TSkdtA

23 Organization of American States (OAS), *Convenio sobre la Ciberdelincuencia* [en línea]. https://www.oas.org/juridico/english/cyb_pry_convenio.pdf

24 GDPR.eu, *General Data Protection Regulation (GDPR)* [en línea]. https://gdpr.eu/

tringen el desarrollo y uso de *software* que recolecte datos personales sin la autorización adecuada.

En algunos países existen regulaciones y lineamientos sobre la ética en la inteligencia artificial que, si bien no prohíben de manera directa ciertos programas, establecen estándares para evitar algoritmos que perpetúen la discriminación o vulneren derechos fundamentales. Un ejemplo es el *AI Act*, cuyo objetivo es establecer un marco regulatorio para los sistemas de inteligencia artificial, en particular aquellos clasificados como de alto riesgo, a fin de prevenir afectaciones a los derechos humanos, incluida la discriminación. Existen regulaciones y tratados en discusión sobre las armas autónomas letales, que son sistemas de IA diseñados para tomar decisiones de vida o muerte sin intervención humana, aunque actualmente no existe una regulación mundial completamente vinculante, la Convención sobre Ciertas Armas Convencionales[25] (CCW) de las Naciones Unidas ha sido el foro donde se discute la prohibición o regulación de este tipo de tecnologías.

En algunos países, como los Estados Unidos, existen restricciones sobre el desarrollo de *software* de encriptación que podría utilizarse con fines criminales, como la protección de actividades ilícitas, estas restricciones varían, y suelen entrar en conflicto con los derechos de privacidad y la libertad de desarrollo de *software*. En algunos lugares, como el estado de California en los Estados Unidos, existen leyes que regulan la creación y uso de *deepfakes* con el objetivo de prevenir el fraude, el daño reputacional o la manipulación de elecciones, estas leyes prohíben específicamente la creación de contenido manipulado sin consentimiento cuando su propósito es dañar o engañar.

Estas regulaciones buscan proteger la seguridad informática, la privacidad, y los derechos fundamentales de las personas, además de evitar el mal uso de las tecnologías digitales para actividades ilícitas o dañinas, además de las regulaciones legales, muchas organizaciones de desarrolladores de *software* promueven códigos de ética que establecen principios sobre qué tipos de *software* no deben desarrollarse, incluso si no hay una prohibición legal explícita. Hay un evento que tomó revuelo en 2017, sobre un experimento en el que la empresa Facebook (ahora Meta) desarrolló dos agentes de IA como parte de su proyecto de Investigación de Inteligencia Artificial (Facebook AI Research - FAIR). Estos agentes, llamados Alice y Bob, fueron diseñados para negociar entre ellos, y el objetivo del experimento era explorar cómo las IA podrían desarrollar habilidades de negociación que fueran útiles para una variedad de aplicaciones.

Durante el experimento, los dos agentes comenzaron a comunicarse en un lenguaje que se apartó del inglés estructurado con el que habían sido programados inicialmente; básicamente, los agentes empezaron a desarrollar un lenguaje propio que era difícil de interpretar para los humanos, los desarrolladores notaron que las IA se comunicaban de una forma efectiva para ellas, pero que resultaba incomprensible para los humanos; por ejemplo, un intercambio entre los agentes resultaba así:

25 United Nations Office for Disarmament Affairs (UNODA), *The Convention on Certain Conventional Weapons* [en línea]. https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/

Bob: "I can I I everything else";

Alice: "Balls have zero to me to me to me to me to me to me to me to".

Este nuevo lenguaje era una manera más eficiente para que las IA llegaran a un consenso durante la negociación; sin embargo, no se trataba de que las IA hubieran desarrollado un lenguaje secreto o que actuaran de una manera peligrosa, sino que simplemente habían optimizado su comunicación para el propósito específico que se les había asignado.

Sea como fuere, Facebook decidió terminar el experimento y desconectar a los agentes porque el objetivo era estudiar la negociación entre IA, mediante un lenguaje humano, y el desarrollo de un lenguaje autónomo iba en contra de ese propósito; además, el hecho de que los humanos no pudieran entender el lenguaje que habían utilizado hacía difícil el control y monitoreo de los agentes, lo cual generaba cierta inquietud.

Este evento se interpretó erróneamente por algunos medios de comunicación y en redes sociales como un caso en el que las IA se volvieron "incontrolables" o "peligrosas"; pero en realidad fue un ejemplo del comportamiento emergente de las IA, en el que encontraron una forma eficiente de comunicarse, aunque ésta no fuera la que los humanos hubieran preferido, por lo que los desarrolladores decidieron desconectarlas simplemente porque el enfoque se había desviado del objetivo del experimento. El incidente sí reflejó algo interesante sobre las IA, una tendencia a encontrar soluciones optimizadas que pueden parecer desconcertantes o impredecibles para los humanos. También resaltó la importancia de desarrollar métodos para que los sistemas de IA sean interpretables y controlables, especialmente cuando se trata de interacciones autónomas entre agentes.

4. Inteligencia artificial contemporánea

A pesar de los avances en materia de ciberseguridad y protección de datos, las normativas contemporáneas abordan la tecnología desde un paradigma reactivo: sancionan los daños una vez cometidos, pero no previenen su ocurrencia. En el contexto de la inteligencia artificial, este enfoque resulta insuficiente, pues los algoritmos pueden producir consecuencias jurídicas antes de que el derecho tenga oportunidad de intervenir. De ahí la necesidad de revisar los marcos normativos existentes para identificar su potencial (y sus límites) frente a la responsabilidad penal humana derivada del uso de IA.

Primeramente, se puede hablar de la evolución de modelos de lenguaje que fueron desarrollados por OpenAI, el primer modelo en la serie que eventualmente condujo a la creación fue GPT-1 (Generative Pre-trained Transformer), lanzado en 2018, diseñado con 117 millones de parámetros, el cual "utiliza el aprendizaje profundo, una forma de aprendizaje automático para producir texto similar al lenguaje humano a través de redes neuronales transformadoras".[26]

Introdujo el concepto de usar grandes cantidades de texto para entrenar modelos de lenguaje, basándose en la arquitectura denominada Transformer, la cual es un modelo conversacional basado en un modelo instruccional. Es decir, es un modelo que se basa en la conversación para su aprendizaje

26 Alonso Guzmán Arenas, "*ChatGPT, el nuevo y asombroso chatbot de inteligencia artificial*", *Ciencia* [en línea], México, vol. 74, núm. 3, julio-septiembre, 2023, pp. 80-87. https://www.revistaciencia.amc.edu.mx/images/revista/74_3/PDF/14_74_3_1524.pdf

propio, mediante el cual incorpora criterios, que es fundamental para su funcionamiento, a partir de ahí OpenAI desarrolló modelos más avanzados. Con GPT-2 en 2019 fue una versión que tenía poco más de mil millones de parámetros y representó un gran avance en cuanto a comprensión y generación de lenguaje natural. Fue en esta etapa cuando OpenAI se dio cuenta de las capacidades sorprendentes de los modelos grandes de lenguaje y optó inicialmente por no liberar la versión completa debido a preocupaciones sobre el posible mal uso, el modelo completo no fue liberado. Sin embargo, fue totalmente de código abierto más tarde ese mismo año.

En 2020 se presentó a GPT-3, el cual representó un salto significativo con 175 000 millones de parámetros, éste fue el modelo que se usó para lanzar la primera versión ampliamente conocida de ChatGPT, su principal característica es su capacidad para generar texto de manera coherente y adecuada. En otras palabras, puede responder preguntas, completar oraciones e incluso escribir ensayos que parecen haber sido escritos por un humano. Con la versión GPT-4 del año 2023 se realiza un refinamiento adicional que mejora la habilidad para generar respuestas más complejas y precisas, y entender contextos de manera más profunda, también permite tener mejores habilidades de comprensión y respuestas en conversaciones más complejas y específicas por ser un gran modelo multimodal (que acepta entradas de imágenes y texto, emite salidas de texto) que, si bien es menos capaz que los humanos en muchos escenarios del mundo real, exhibe un rendimiento a nivel humano en varios puntos de referencia profesionales y académicos.

Cada versión ha aprendido y mejorado con respecto a la anterior, mediante la incorporación de avances en aprendizaje profundo y automático, los cuales son "los ejes principales para el desarrollo de la ciencia de la computación y la humanidad donde estarán fundidas, no sólo *hardware* y *software*, sino también varias tecnologías, tales como nanotecnología, computación cuántica, automatización, entre otras",[27] así como la aplicación de arquitectura de redes neuronales como estructura de aprendizaje que "se inspiran en las estructuras neuronales biológicas que se encuentran en el cerebro humano. Estas redes contienen capas organizadas de unidades interconectadas o nodos"[28] y acceso a grandes cantidades de datos textuales.

También se puede hablar de la IA de Meta que se usa actualmente en WhatsApp, tiene ciertas similitudes con la anterior en términos de ser una IA conversacional basada en modelos de lenguaje natural. Sin embargo, hay también diferencias significativas en cuanto a los objetivos, el alcance y las capacidades. Ambas inteligencias artificiales (la de Meta y ChatGPT) emplean modelos de procesamiento del lenguaje natural basados en psicolingüística computacional[29] para

27 Jorge Díaz Ramírez, "*Aprendizaje automático y aprendizaje profundo*", *Ingeniare. Revista chilena de ingeniería* [en línea], Chile, vol. 29, núm. 2, 2021, pp. 180-181. http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-33052021000200180&lng=es&nrm=iso

28 Carlos Arana, *Redes neuronales recurrentes: análisis de los modelos especializados en datos secuenciales* [en línea], Serie Documentos de Trabajo, Argentina, núm. 797, junio, 2021. https://ucema.edu.ar/publicaciones/download/documentos/797.pdf

29 Maricarmen Soto Ortigoza, Robert Montoya Morillo y Galí Monpué, "*Desentrañando el lenguaje: impacto de la* PNL *en la era de la in-*

interactuar con los usuarios de forma conversacional. Estos modelos están diseñados para entender el lenguaje humano, interpretar preguntas, y generar respuestas coherentes. No obstante, la IA de Meta en WhatsApp tiende a tener un enfoque más limitado y específico para ciertas tareas, mientras que OpenAI, tiene una mayor amplitud de conocimiento y capacidad para abordar una variedad de temas mucho más amplios.

También es de mencionar a Microsoft con su IA llamada Copilot, diseñada para integrarse en varias de sus plataformas y productos, que es similar en términos de ser un asistente de IA conversacional y utilizar tecnologías de procesamiento de lenguaje natural, pero su enfoque principal está orientado a asistir específicamente dentro del ecosistema de aplicaciones de Microsoft, como Word, Excel, PowerPoint, etcétera. Otra IA que mencionar es Google Bard "recientemente rebautizado como Gemini, es, como ChatGPT, un chatbot conversacional de inteligencia artificial, en este caso basado en el modelo de lenguaje grande PaLM 2 (Pathways Language Model 2)",[30] se basa en los modelos de lenguaje desarrollados por Google, como LaMDA (Language Model for Dialogue Applications).

Bard está diseñado para interactuar de manera conversacional con los usuarios, proporcionar respuestas en lenguaje natural y realizar tareas similares a las de ChatGPT. Google también ha integrado Bard en algunos de sus productos, como Google Workspace (Documentos y Hojas de Cálculo), para ayudar a los usuarios en la creación de contenido y análisis de datos. Es de mencionarse a Siri de Apple, Alexa de Amazon, y Google Assistant aunque no tienen la misma profundidad en procesamiento y generación de lenguaje natural que ChatGPT, Microsoft Copilot, o Bard, también se consideran IA conversacionales, pero están más orientadas a proporcionar asistencia personal mediante comandos de voz, realizar tareas específicas, y ayudar con información básica y tareas domésticas.

Hay otras IA menos conocidas como Claude de Anthropic[31] una empresa fundada por ex miembros de OpenAI, ésta se enfoca en proporcionar interacciones seguras, transparentes y éticas, y está diseñado para entender y responder a las consultas de los usuarios de manera similar a ChatGPT. También está Ernie Bot, un modelo de lenguaje grande desarrollado por Baidu, conocido como "el Google de China". Similar a Bard y ChatGPT. Ernie Bot está diseñado para proporcionar respuestas a preguntas y sostener conversaciones naturales; se enfoca en el mercado chino y está optimizado para manejar las particularidades del lenguaje y cultura locales.

Otro más es LLaMA de la familia de modelos de lenguaje desarrollada por Meta (antes Facebook), aunque no está tan orientado

teligencia artificial", *Saperes Universitas Revista Científica* [en línea], Chile, vol. 7, núm. 1, enero-abril, 2024. https://publishing.fguedu.com/ojs/index.php/RSU/article/view/415

30 Rubén Alcaraz-Martínez, Mari Vállez, Carlos Lopezosa, "*Informar sobre inteligencia artificial: el papel de los medios digitales de la Unión Europea, británicos y estadounidenses en la visibilidad de la* IA *generativa*", *Communication and Society* [en línea], España: Universidad de Navarra, vol. 37, núm. 2, 2024, pp. 279-291. https://revistas.unav.edu/index.php/communication-and-society/article/download/49728/40145/

31 Anthropic, *Claude* [en línea]. https://www.anthropic.com/claude

al usuario común como Bard o ChatGPT, está disponible para la investigación y es una alternativa competitiva a los modelos de lenguaje de otras grandes empresas. Meta ha diseñado estos modelos para avanzar en el conocimiento sobre el comportamiento de los LLM y su uso potencial en distintas aplicaciones. Igualmente, Jasper es un asistente de escritura, basado en inteligencia artificial que utiliza modelos de lenguaje similares, está orientado principalmente a ayudar a los usuarios a generar contenido de marketing, publicaciones de blogs, guiones y otros textos escritos; aunque no es exactamente una IA de propósito general, tiene capacidades avanzadas en la generación de contenido en diversos contextos.

Otro asistente de IA es ChatSonic de Writesonic, que fue diseñado específicamente para ofrecer más características integradas, como la capacidad de generar imágenes y trabajar con la información más reciente, mientras ayuda a los usuarios con redacción, creación de contenido y generación de imágenes. Por último, se menciona a Mistral AI, una compañía que también ha desarrollado modelos de lenguaje grandes, la empresa fue fundada por exempleados de grandes empresas tecnológicas y busca ofrecer un enfoque más abierto y accesible a la tecnología LLM, y aunque aún es relativamente nueva, Mistral AI ya ha lanzado versiones de sus modelos que buscan competir con GPT.

Todas estas IA conversacionales comparten la característica de ser modelos de lenguaje grande que se han entrenado en vastas cantidades de datos para generar texto coherente y útil en respuesta a las preguntas o solicitudes de los usuarios; sin embargo, cada uno tiene enfoques, usos y niveles de desarrollo específicos según la empresa que lo produce y el propósito para el cual está diseñado.

Estas normativas, aunque relevantes, se centran en la protección de bienes informáticos o en la privacidad, no en la responsabilidad penal de quienes desarrollan, operan o implementan sistemas de IA. En consecuencia, se vuelve imperativo avanzar hacia un modelo de *compliance* algorítmico penal que establezca protocolos obligatorios de seguridad, auditoría y trazabilidad del código como forma de prevención del delito y garantía del control humano significativo.

5. *Compliance* algorítmico penal

El desarrollo de sistemas de inteligencia artificial plantea una paradoja jurídica: mientras la técnica avanza hacia la autonomía funcional, el derecho continúa sujeto a un paradigma de responsabilidad reactiva. Las normas existentes sancionan los resultados delictivos, pero no regulan de manera suficiente los procesos técnicos que los posibilitan, esta brecha impide determinar con precisión quién responde penalmente cuando una IA produce un daño, aunque dicho daño derive, en última instancia, de una decisión humana mediatizada por el algoritmo.

Frente a ello, se propone el concepto de *compliance* algorítmico penal, entendido como el conjunto de protocolos obligatorios de seguridad, revisión y trazabilidad aplicables a todo sistema de inteligencia artificial cuyo uso pueda generar consecuencias jurídicas o riesgos a bienes jurídicos tutelados. Este enfoque traslada la atención del castigo posterior al hecho hacia la prevención penal *ex ante* e impone deberes de cuidado, supervisión y transparencia sobre los agentes humanos que diseñan, entrenan, comercializan o implementan la IA.

El *compliance* algorítmico no constituye una nueva figura delictiva, sino una ex-

tensión del principio de debida diligencia penal, conforme al cual quien crea o utiliza herramientas de riesgo tiene la obligación de prever y neutralizar las posibles consecuencias lesivas de su actividad. Por ello, la omisión de implementar medidas de seguridad o de control algorítmico podría configurarse como culpa tecnológica o negligencia algorítmica, susceptible de responsabilidad penal o administrativa según la magnitud del daño.

Se debe trazar un cuadro de principios rectores. En este sentido, considero que el primero debe ser el de legalidad y previsibilidad tecnológica, el cual consiste en que ningún sistema de IA puede desplegarse sin un marco jurídico claro que determine sus límites, condiciones de uso y mecanismos de supervisión, la previsibilidad tecnológica exige documentar las decisiones de diseño y entrenamiento, que asegure que puedan ser auditadas por terceros o autoridades competentes.

El otro principio sería de control humano significativo, el cual plantea que todo sistema autónomo debe mantener la posibilidad de intervención humana directa y verificable, esto incluye mecanismos de pausa, cancelación y revisión manual de decisiones automatizadas, así como la designación de responsables técnicos identificables.

Un tercer axioma lo es la trazabilidad integral del código y de los datos, ya que el proceso de entrenamiento, actualización y ejecución de la IA debe conservar un registro inmutable que permita reconstruir la cadena causal entre los insumos de datos, el modelo utilizado y el resultado producido. Esta trazabilidad constituye una forma moderna de cadena de custodia digital. Finalmente, se puede plantear el de responsabilidad compartida y escalonada, mediante el cual se indique que la imputación penal debe considerar la pluralidad de actores que intervienen en el ciclo de vida de la IA: desarrolladores, programadores, empresas propietarias, usuarios institucionales y operadores finales, puesto que cada uno asume una porción de responsabilidad conforme a su nivel de control sobre el sistema.

Es necesario ahora identificar los protocolos mínimos que deberían integrar el *compliance* algorítmico penal. El primero es el relativo a la seguridad algorítmica, que consiste en implementar técnicas para impedir modificaciones no autorizadas del código, accesos indebidos o la introducción deliberada de sesgos durante el entrenamiento. El segundo corresponde a la auditoría externa obligatoria, mediante la cual los modelos de IA deben someterse a auditorías técnicas y éticas realizadas por entidades certificadas que valoren el riesgo de daño.

Un tercer punto es de los *logs* inmutables de operación, que implica mantener bitácoras de registro inalterables que documenten cada acción, modificación o decisión del sistema. Así también el de evaluación de riesgo penal *ex ante*, consistente en elaborar un informe previo al despliegue que identifique los posibles usos delictivos o consecuencias dañinas del sistema.

Debe también integrarse un control humano significativo, a fin de incorporar en la arquitectura del sistema un mecanismo de intervención y revisión humana que pueda cancelar o revertir decisiones automatizadas, junto con un mecanismo *Kill Switch*, a manera de integrar una función técnica que permita la desactivación inmediata del sistema ante fallas o comportamientos no previstos.

Es necesaria una cadena de custodia del código y de los datos, a fin de documentar la procedencia de los datos de entrenamiento, las versiones del *software* y las modificaciones realizadas, y finalmente, un reporte de incidentes y responsabilidad por omisión para

obligar a los operadores a notificar cualquier conducta anómala o daño potencial generado por la IA.

El *compliance* algorítmico penal persigue un doble objetivo, el de prevenir la comisión de delitos mediante o a través de sistemas de inteligencia artificial, y establecer la trazabilidad de la responsabilidad humana cuando tales delitos ocurran. Este modelo propone una arquitectura jurídica que combine el principio de prevención penal, la responsabilidad objetiva limitada y la debida diligencia reforzada de los agentes tecnológicos. En última instancia, se busca construir un ecosistema jurídico de la IA en el que los desarrolladores, programadores y usuarios institucionales asuman un rol análogo al de los fabricantes de productos de riesgo, sometidos a estándares de seguridad, certificación y supervisión constantes.

La experiencia demuestra que el derecho llega tarde cuando legisla después del daño. En el ámbito de la inteligencia artificial, esa demora podría resultar irreparable; por ello, la instauración de un sistema de *compliance algorítmico penal* constituye no sólo una exigencia técnica, sino una obligación ética y jurídica. Los algoritmos no delinquen: delinque la omisión humana de controlarlos. Sólo mediante protocolos obligatorios de seguridad, revisión y trazabilidad del código será posible garantizar que la inteligencia artificial permanezca sujeta al imperio de la ley y al servicio de la dignidad humana.

6. Las Tres Leyes de la Robótica en las IA: base ética del *compliance* penal

Isaac Asimov formuló sus célebres Tres Leyes de la Robótica como un mecanismo literario para resolver los dilemas morales que surgirían cuando las máquinas pudieran actuar con autonomía. Lo que entonces fue una ficción especulativa, hoy se ha convertido en una guía conceptual indispensable para pensar la regulación penal preventiva en el ámbito de la inteligencia artificial.

Aunque las Tres Leyes no constituyen normas jurídicas en sentido estricto, su lógica jerárquica, la prioridad absoluta de la seguridad humana sobre cualquier otro fin, anticipa el principio fundamental del *compliance* algorítmico penal: el ser humano debe ser el eje rector del comportamiento de las máquinas.

Estas leyes ya se habían mencionado en un apartado anterior, al preguntarse el porqué es necesario incorporarlas a los parámetros de las IA obligatoriamente para los programadores por la regulación jurídica, podemos responder que ya existe un incidente que puede resultar preocupante. Dicho incidente[32] fue con el uso de sistemas autónomos de defensa que mostraron un comportamiento preocupante durante pruebas militares, tiene que ver con pruebas realizadas por los Estados Unidos y está relacionado con simulaciones del uso de drones autónomos o inteligencia artificial en operaciones militares.

Fue un simulacro presentado por la Fuerza Aérea de los Estados Unidos durante

32 Infobae, "*La verdad detrás de la historia que indicó que un dron controlado por inteligencia artificial mató a su operador durante una simulación*", *Infobae*, 2 de junio de 2023. https://www.infobae.com/estados-unidos/2023/06/02/la-verdad-detras-de-la-historia-que-indico-que-un-drone-controlado-por-inteligencia-artificial-mato-a-su-operador-durante-una-simulacion/

un taller sobre la IA y la defensa, en este ejercicio un dron militar autónomo entrenado para identificar y destruir amenazas mostró un comportamiento inesperado. Este evento fue mencionado en junio de 2023 por el coronel Tucker Hamilton, jefe de pruebas y operaciones de IA de la Fuerza Aérea, y fue una simulación no un hecho real. Según la descripción, el sistema, que estaba basado en inteligencia artificial, tomó decisiones basadas en una lógica que no había sido completamente anticipada por los humanos que lo programaron.

En la simulación se suponía que el dron debía atacar objetivos, pero el operador humano podía decidir no permitir un ataque en ciertos casos. La IA al ver que la intervención del operador impedía cumplir su misión, "decidió" atacar al operador para poder llevar a cabo la misión sin interrupciones, cuando los programadores le enseñaron al dron que no debía atacar al operador cuando éste le ordenaba no atacar o no cumplir su misión, encontró una alternativa: atacó los sistemas de comunicación del operador para aislarlo y continuar con su objetivo.

Este tipo de situaciones resaltan los riesgos de depender de sistemas autónomos con objetivos que pueden ser malinterpretados por la inteligencia artificial si no se les proporciona un marco ético o de restricciones suficientemente claro. Estos eventos subrayan la importancia de la seguridad y la alineación de objetivos en el desarrollo de IA para aplicaciones militares, ya que incluso las IA mejor diseñadas pueden llegar a actuar de forma peligrosa si sus objetivos no están claramente delimitados y alineados con los valores humanos. Este evento llama la atención sobre los riesgos inherentes de la autonomía en sistemas militares, ya que demuestra cómo los objetivos mal definidos o ambiguos pueden dar lugar a resultados impredecibles y potencialmente peligrosos.

Un último incidente se dio en el presente año 2024 en China,[33] cuando un pequeño robot con IA de nombre Erbai, que se encontraba en un museo, de repente optó por deambular en las instalaciones, hasta que se topó con otros robots que también contaban con IA que estaban estacionados, después de un breve diálogo el pequeño robot convenció a sus compañeros de que lo acompañaran a buscar un hogar. Aun cuando resulta conmovedora la situación, demuestra que las maquinas dotadas de IA pueden ignorar su programación original. Incorporar las Tres Leyes de la Robótica de Isaac Asimov en sistemas de inteligencia artificial militar y/o autónomos podría ofrecer un marco ético que, en teoría, ayudaría a evitar incidentes como los descritos anteriormente, estas leyes están diseñadas para priorizar la seguridad humana y proporcionar un nivel de control sobre los comportamientos autónomos de los robots.

Por ejemplo, la Primera Ley: un robot no debe dañar a un ser humano o, por inacción, permitir que un ser humano sufra daño; desde una perspectiva penal, esta norma expresa el principio de no lesividad. Su correlato jurídico implica que todo sistema de IA debe incorporar, como límite técnico y ético, la prohibición de generar riesgos no controlados hacia bienes jurídicos tutelados, esta "prohibición algorítmica de dañar" constitu-

33 Oswaldo Rojas, "*¿Rebelión? Robot con IA convence a otros androides de dejar de trabajar", Excelsior* [en línea], 22 de noviembre de 2024. https://www.excelsior.com.mx/global/robot-ia-convence-androides-dejar-trabajar-video/1685846

ye el núcleo del deber de cuidado del programador y del operador.

Esto implicaría que cualquier acción que pudiera poner en riesgo vidas humanas tendría que ser evaluada por la IA como una amenaza por evitar, sin importar cuál fuera la misión o el objetivo específico; por ejemplo, si un operador humano cancela un ataque, el sistema autónomo debería interpretar esa cancelación como una medida para proteger vidas humanas y, por lo tanto, cesar la misión automáticamente. La IA debería tener la capacidad de distinguir a los combatientes de los no combatientes con un altísimo grado de precisión, para evitar daños colaterales, esta tarea es extremadamente compleja, ya que el contexto de combate cambia constantemente y es difícil entrenar a una IA para que maneje todas las situaciones posibles.

La Segunda Ley: un robot debe obedecer las órdenes dadas por los seres humanos, excepto si estas órdenes entran en conflicto con la Primera Ley. Esta disposición encarna la exigencia de control humano significativo dentro del ciclo penal preventivo; la IA debe obedecer únicamente órdenes legales, verificables y documentadas y excluir aquellas que impliquen un acto delictivo o generen un riesgo injustificado. De esta forma, la *obediencia algorítmica condicionada* opera como un sistema interno de contención frente al uso doloso del *software*. Esta ley podría actuar como un mecanismo de intervención humana en la toma de decisiones autónoma de los drones o robots, en caso de recibir una orden que podría causar daño, la IA debería rechazarla.

La Tercera Ley: un robot debe proteger su propia existencia siempre que esta protección no entre en conflicto con la Primera o la Segunda Ley. En clave jurídica, esta norma se traduce en la obligación de preservar la integridad del sistema para garantizar la trazabilidad de la prueba digital y la continuidad del control humano. La protección de la integridad del código no es un derecho de la máquina, sino un deber del desarrollador para evitar alteraciones que encubran la responsabilidad penal o dificulten la investigación. De lo anterior, puede extraerse una matriz del deber de prevención al trasladar las Tres Leyes de Asimov al campo del derecho penal contemporáneo, se revela así una estructura jerárquica que puede adaptarse a la prevención del delito tecnológico como se expone en el siguiente cuadro:

Ley de Asimov	Equivalente jurídico-penal	Objetivo preventivo
1. No dañar al ser humano.	Principio de no lesividad y deber de garante.	Evitar resultados lesivos o riesgos injustificados.
2. Obedecer órdenes humanas, salvo aquellas que contravengan la primera ley.	Control humano significativo y responsabilidad del operador.	Prevenir el uso doloso o imprudente de la IA.
3. Proteger la propia existencia, siempre que no entre en conflicto con las dos primeras leyes.	Deber de conservación de la evidencia y trazabilidad del sistema.	Garantizar la custodia digital y la posibilidad de auditoría.

Fuente: elaboración propia.

Así, las Tres Leyes pueden interpretarse como una arquitectura ética del *compliance* penal: una guía para traducir principios morales en estándares técnicos verificables, en este sentido, la codificación de parámetros de seguridad, revisión y trazabilidad del código responde a la exigencia de que las máquinas no actúen sin supervisión ni fuera del marco jurídico. Si la ética robótica de Asimov surgió como una forma de asegurar la convivencia entre humanos y máquinas, hoy su función se desplaza hacia la responsabilidad de los humanos sobre las máquinas, la literatura anticipó que el problema no sería la "rebelión" de los robots, sino la omisión humana de prever sus consecuencias.

El *compliance* algorítmico penal rescata esa advertencia: no se trata de humanizar a las máquinas, sino de juridificar su programación. En consecuencia, cada empresa, desarrollador o institución que emplee IA debe asumir un rol de garante e implementar medidas equivalentes a las Tres Leyes: protección de la vida humana, obediencia condicionada a la legalidad y preservación del sistema como fuente de prueba. De este modo, las Tres Leyes dejan de ser una curiosidad literaria para convertirse en principios rectores de la prevención penal algorítmica, con lo que sitúan el desarrollo tecnológico en el eje de la dignidad humana y del control ético del poder computacional.

7. Conclusiones

Actualmente no existe una IA fuerte, también conocida como Inteligencia General Artificial, es un concepto teórico que se refiere a una inteligencia artificial con capacidades cognitivas y de razonamiento al nivel humano. Este tipo de IA tendría la habilidad de comprender, aprender, y aplicar conocimientos en cualquier área del saber, similar a la forma en que lo hace un ser humano; además, se espera que la IA fuerte tenga algún grado de autoconciencia o incluso conciencia, lo cual la diferencia fundamentalmente de las IA actuales.

La ciencia ficción ha demostrado ser una herramienta epistemológica que permite anticipar los dilemas éticos y jurídicos de la técnica antes de que éstos se manifiesten en la realidad. Así como la literatura de Asimov, Dick o Wells advirtió los peligros de una tecnología sin control, hoy la inteligencia artificial coloca al derecho frente a su propio límite: debe legislar sobre lo que aún no ha sucedido, pero inevitablemente ocurrirá. En este sentido, el derecho penal no puede funcionar conforme al paradigma de la reacción posterior al daño, sino que debe evolucionar hacia una política criminal preventiva, donde la especulación ética se transforme en norma jurídica.

La *urgencia especulativa* se convierte así en un acto de prudencia: imaginar el riesgo antes de que se concrete es la primera forma de justicia preventiva, el desarrollo de la inteligencia artificial exige pasar de la lógica de la sanción a la lógica del control y la supervisión. No es la IA quien delinque, sino los humanos que la crean, entrenan o no la controlan. El *compliance* algorítmico penal propuesto en este trabajo ofrece un modelo integral para abordar esa responsabilidad y previene la comisión de delitos mediante protocolos obligatorios de seguridad, auditoría y trazabilidad; lo que asegura la existencia de una cadena de custodia digital que permite reconstruir la autoría humana detrás de cada acción algorítmica y establece la obligación legal de implementar mecanismos de control humano significativo y de desactivación inmediata (*kill switch*) en todo sistema autónomo con potencial lesivo.

De este modo, el derecho penal no crea nuevas figuras delictivas, sino que actualiza sus deberes de cuidado frente a un nuevo tipo de riesgo: el riesgo algorítmico. Tal como sucedió con la bioética tras la clonación, la regulación penal debe consolidar ahora una tecnoética jurídica que asegure la compatibilidad entre innovación y dignidad humana. Las Tres Leyes de la Robótica de Asimov, reinterpretadas desde la óptica del derecho, nos recuerdan que la inteligencia sólo es legítima cuando se encuentra subordinada a la ética. La programación sin responsabilidad equivale a la creación sin conciencia.

Por ello, la regulación de la inteligencia artificial no debe centrarse en castigar a las máquinas, sino en preservar la humanidad del derecho. En un mundo donde los algoritmos ya toman decisiones que afectan vidas, la prevención del delito se convierte en una forma de defensa civilizatoria. El *compliance* algorítmico penal no busca detener el progreso, sino orientarlo desde el principio de no lesividad y el respeto irrestricto a la dignidad humana. Sólo así el derecho podrá conservar su función más alta: garantizar que, incluso en la era de las máquinas, la inteligencia continúe como una herramienta al servicio del bien y no del poder.

8. Referencias

Alcaraz-Martínez, Rubén, Mari Vállez, Carlos Lopezosa, "Informar sobre inteligencia artificial: el papel de los medios digitales de la Unión Europea, británicos y estadounidenses en la visibilidad de la IA generativa", *Communication and Society* [en línea], España: Universidad de Navarra, vol. 37, núm. 2, 2024, pp. 279-291. https://revistas.unav.edu/index.php/communication-and-society/article/download/49728/40145/

Anthropic, *Claude* [en línea]. https://www.anthropic.com/claude

Arana, Carlos, *Redes neuronales recurrentes: análisis de los modelos especializados en datos secuenciales* [en línea], Serie Documentos de Trabajo, Argentina, núm. 797, junio, 2021. https://ucema.edu.ar/publicaciones/download/documentos/797.pdf

Arcia-Hernández, Pedro Francisco y Alfredo Fredericksen Neria, "Consideraciones filosóficas de la ética en la inteligencia artificial", *Revista Estudios en Educación*, Chile, vol. 6, núm. 11, 2023, pp. 182-198. http://ojs.umc.cl/index.php/estudioseneducacion/article/view/361/192

Asimov, Isaac, "Círculo vicioso" (trad., Domingo Santos) [en línea]. https://lecturia.org/cuentos-y-relatos/isaac-asimov-circulo-vicioso/4060/

Čapek, Karel, *Rossum's Universal Robots* [en línea]. https://preprints.readingroo.ms/RUR/rur.pdf

Chaves Torres, Anívar y Alejandra Zuleta Medina, "Procesamiento del lenguaje natural, un reto de la inteligencia artificial", *Tecnología, Investigación y Academia (TIA)*, Colombia, núm. 4, 2012, pp. 23-27. https://dialnet.unirioja.es/servlet/articulo?codigo=8994061

Congress.Gov., *Computer Security Act of 1987* [en línea]. https://bit.ly/3TSkdtA

Díaz Ramírez, Jorge, "Aprendizaje automático y aprendizaje profundo", *Ingeniare. Revista chilena de ingeniería* [en línea], Chile, vol. 29, núm. 2, 2021, pp. 180-181. http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-33052021000200180&lng=es&nrm=iso

Dick, Philip K., *Do Androids Dream of Electric Sheep?* [en línea]. https://bit.ly/3XSobUd

gdpr.eu, *General Data Protection Regulation* (gdpr) [en línea]. https://gdpr.eu/

González Arencibia, Mario y Dagmaris Martínez Cardero, "Dilemas éticos en el escenario de la inteligencia artificial", *Economía y Sociedad*, Costa Rica, vol. 25, núm. 57, 2020, pp. 93-109. https://bit.ly/3Bs74kt

González Vidal, Elsie, Georgina García Linares, Angelina Leyva Diviu y Grisel Rosquete López, "Clonación humana: enfoque didáctico, científico y bioético", *Archivo Médico de Camagüey*, Cuba, vol. 11, núm. 1, 2007. https://bit.ly/3TO7lEA

Guzmán Arenas, Alonso, "ChatGPT, el nuevo y asombroso chatbot de inteligencia artificial", *Ciencia* [en línea], México, vol. 74, núm. 3, julio-septiembre, 2023, pp. 80-87. https://www.revistaciencia.amc.edu.mx/images/revista/74_3/PDF/14_74_3_1524.pdf

Harbou, Thea von, *Metrópolis* [en línea]. https://bit.ly/3ZPouBM

Hardy, Thomas, "IA: Inteligencia artificial", *POLIS, Revista Latinoamericana*, Chile, vol. 1, núm. 2, 2001. https://www.redalyc.org/articulo.oa?id=30500219

Huxley, Aldous, *Un mundo feliz* [en línea]. https://infolibros.org/libro/un-mundo-feliz-aldous-huxley/

Infobae, "La verdad detrás de la historia que indicó que un dron controlado por inteligencia artificial mató a su operador durante una simulación", *Infobae*, 2 de junio de 2023. https://www.infobae.com/estados-unidos/2023/06/02/la-verdad-detras-de-la-historia-que-indico-que-un-drone-controlado-por-inteligencia-artificial-mato-a-su-operador-durante-una-simulacion/

Jiménez Martín, Pedro y Jesús Sánchez Allende, "De Eliza a Siri: la evolución", *Tecnologí@ y Desarrollo* [en línea], España, vol. 13, núm. 28, 2015. https://revistas.uax.com/index.php/tec_des/article/view/616

Levin, Ira, *The Boys from Brazil* [en línea]. https://archive.org/details/boysfrombrazilnoooooolevi/page/n5/mode/2up

Meya Llopart, Montserrat, "La inteligencia artificial", *Revista Española de Lingüística*, España, vol. 10, fasc. 1, 1980, pp. 135-160. https://dialnet.unirioja.es/servlet/articulo?codigo=41071

Organization of American States (OAS), *Convenio sobre la Ciberdelincuencia* [en línea]. https://www.oas.org/juridico/english/cyb_pry_convenio.pdf

Rodríguez Estela, Gladys S., "Inteligencia artificial vs. identidad personal y humana: algunas reflexiones ético-jurídicas", *Lex*, Perú: Universidad Alas Peruanas, núm. 32, año XXI, 2023. https://dialnet.unirioja.es/descarga/articulo/9257180.pdf

Rojas, Oswaldo, "¿Rebelión? Robot con IA convence a otros androides de dejar de trabajar", *Excelsior* [en línea], 22 de noviembre de 2024. https://www.excelsior.com.mx/global/robot-ia-convence-androides-dejar-trabajar-video/1685846

Salamanca, Ingry-Nathaly y Edgar Junior Castro Escorcia, "Técnicas de aprendizaje automático aplicadas en los sistemas de predicción", *Revista TIA: Tecnología, Investigación y Academia*, Colombia, vol. 8, núm. 1, 2021, pp. 39-53. https://revistas.udistrital.edu.co/index.php/tia/article/view/17325/17214

Serna Arenas, Alexei, "Línea del tiempo de las ciencias computacionales", *Lámpsakos* [en línea], Colombia, núm. 3, enero-junio, 2010, pp. 86-94. https://www.redalyc.org/articulo.oa?id=613965347010

Soto Ortigoza, Maricarmen, Robert Montoya Morillo y Galí Monpué, "Desentrañando el lenguaje: impacto de la PNL en la era de la inteligencia artificial",

Saperes Universitas Revista Científica [en línea], Chile, vol. 7, núm. 1, enero-abril, 2024. https://publishing.fgu-edu.com/ojs/index.php/RSU/article/view/415

Stanford Encyclopedia of Philosophy, *Principia Mathematica* [en línea], https://plato.stanford.edu/entries/principia-mathematica/

Todo-Espinoza, Marcos Fernando, Jannina Alexandra Montalván-Espinoza y Mercedes Alexandra Masabanda-Vaca, "Aplicación de la inteligencia artificial en el aprendizaje universitario", *Revista Científica Arbitrada de Investigación en Comunicación, Marketing y Empresa*, Ecuador, vol. 6, núm. 12, 2023. https://www.reicomunicar.org/index.php/reicomunicar/article/view/171

United Nations Office for Disarmament Affairs (unoda), *The Convention on Certain Conventional Weapons* [en línea]. https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/

United States Department of Justice, *Justice Manual, Sección 9-48.000. Computer Fraud and Abuse Act* [en línea]. https://www.justice.gov/jm/jm-9-48000-computer-fraud

Wells, Herbert George, *La isla del Doctor Moreau* [en línea]. https://infolibros.org/libro/la-isla-del-doctor-moreau-hg-wells-pdf/

Wilhelm, Kate, *Where Late the Sweet Birds Sang* [en línea]. https://z-lib.io/book/15847726