Seminario Ética e Inteligencia Artificial

Implicaciones éticas derivadas del impacto de la inteligencia artificial en la sociedad contemporánea

Juan Pablo Arias Remolina, Yonathan Camilo Benítez Mancipe, Andrés Felipe Cárdenas Zárate, Sonia Marcela Granados Moreno y Neyder Fabian Mosquera Niño

Trabajo de Grado para Optar al Título de Ingeniero de Sistemas

Modalidad

Seminario de Investigación

Director

Sonia Cristina Gamboa Sarmiento

Doctora en Educación

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingeniería de Sistemas e Informática

Ingeniería de Sistemas

Bucaramanga

2025

Dedicatoria

Dedico esta tesis a mi familia, fuente de amor, valentía e inspiración; a mí mismo, por la constancia y el esfuerzo durante toda la carrera; a la UIS, por brindarme un espacio de crecimiento académico y personal; y al pueblo colombiano que a través de sus impuestos me dieron la oportunidad de estudiar un pregrado en una universidad pública de calidad.

Juan Pablo Arias Remolina

A Dios, por permitirme llegar a este momento tan importante de mi vida. También agradezco la confianza y el apoyo de mis padres quienes con sus consejos han sabido guiarme para culminar mi carrera profesional. A mis hermanas que me han dado un impulso emocional, siendo mi refugio en los días difíciles y mi mayor fuente de motivación. Finalmente, a la profesora Sonia Gamboa, gracias por su tiempo, por su apoyo, así como por la sabiduría que nos transmitió durante este seminario y parte de la carrera profesional.

Yonathan Camilo Benítez Mancipe

Primeramente, quiero agradecer a Dios por permitir que todo el esfuerzo de una carrera se haya podido plasmar en este libro, tantas noches en vela, tanto esfuerzo y dedicación que finalmente están dando sus frutos. A mis padres, porque gracias a ellos estoy en la universidad, porque fueron mi guía, mi consejo y mi espacio de tranquilidad. A mi novia, por apoyarme incondicionalmente, por su tiempo, apoyo y su fuerza trasmitida a través del amor. A mi hermanito por ser la risa en medio de la adversidad, una mirada a mi niñez y fuente de alegría. A la docente Sonia Gamboa por su disposición, por su carisma, por nunca frenar frente a los obstáculos y por haber sido la mejor guía que este proyecto pudo haber tenido.

Andrés Felipe Cárdenas Zárate

A Dios, por ser mi guía, por darme la fortaleza en los momentos difíciles y por acompañarme en cada paso de este camino. A mis padres por su apoyo incondicional, por ser mi mayor ejemplo de constancia, esfuerzo y entrega, su amor es una de las razones por las que hoy estoy aquí. A mi abuela por ser mi refugio y a sus oraciones que siempre me acompañaron. Finalmente, a la profesora Sonia Gamboa, gracias por su tiempo, por guiarnos con paciencia y compromiso, por compartir su conocimiento a lo largo de este seminario y durante nuestra formación académica.

Sonia Marcela Granados Moreno

A mi madre, Esmeralda Niño, por ser esa fuerza silenciosa y motivadora. Tu amor fue empuje, escudo y hogar.

A mi padre, Gustavo Mosquera, por enseñarme que la disciplina y la rectitud abren caminos cuando todo parece estar en contra.

A mis hermanos, Cristian y Gustavo, por ser compañía, risa, caos y apoyo incondicional en cada etapa del camino.

Y a mi gato, Volibear, mi compañero fiel durante tantas noches largas y silenciosas. Por ser la sombra que acompañó mis desvelos, el eco que respondía a mi alma en los momentos de soledad. En su mirada hallé consuelo; en su presencia, la calma que necesitaba para seguir adelante. Gracias por esos ronroneos que, sin decir nada, lo decían todo.

Neyder Fabian Mosquera Niño

Agradecimientos

A lo largo del desarrollo de este proyecto, hemos contado con el acompañamiento, apoyo y guía de muchas personas a quienes hoy queremos expresar nuestra más sincera gratitud.

En primer lugar, agradecemos a nuestros familiares, su respaldo ha sido el pilar fundamental que nos ha sostenido durante este proceso académico y personal.

A nuestra directora, PhD. Sonia Cristina Gamboa Sarmiento, cuyo compromiso y disposición fueron clave para el desarrollo exitoso de este seminario.

A nuestros compañeros y amigos, gracias por los momentos compartidos, las ideas discutidas, los apoyos desinteresados y la compañía en el recorrido.

Finalmente, agradecemos a la institución que nos formó, por brindarnos los recursos, espacios y oportunidades para crecer como profesionales y como personas.

Este logro no es solo nuestro; también pertenece a todos aquellos que creyeron en nosotros y caminaron, de una u otra manera, a nuestro lado.

Juan Pablo Arias Remolina

Yonathan Camilo Benítez Mancipe

Andrés Felipe Cárdenas Zárate

Sonia Marcela Granados Moreno

Neyder Fabian Mosquera Niño

Tabla de Contenido

Pág.

Lista de Tablas

Pág.

Lista de Figuras

Pág.

Lista de Apéndices

pág.

Resumen

Título: Seminario de Investigación Ética e Inteligencia Artificial: implicaciones éticas derivadas del impacto de la inteligencia artificial en la sociedad contemporánea*

Autor: Juan Pablo Arias Remolina
Yonathan Camilo Benítez Mancipe
Andrés Felipe Cárdenas Zárate
Sonia Marcela Granados Moreno
Neyder Fabian Mosquera Niño**

Palabras Clave: Inteligencia artificial, ética en inteligencia artificial, inteligencia artificial general, responsabilidad tecnológica, regulación de inteligencia artificial.

Descripción:

En la actualidad, la computación, cuyo interés se centra, principalmente, en desarrollar hardware y software eficiente que logre ejecutar tareas cada vez más complejas, y que van desde la imitación y superación de tareas típicamente humanas hasta procesos que las limitaciones biológicas humanas no les permiten realizarlos sin ayuda tecnológica. Si bien, los beneficios en todas las áreas de la vida, de la industria, de la ciencia, derivados de la computación son incalculables, hoy en día, con el auge de la Inteligencia artificial, la humanidad se enfrenta a dilemas, riesgos, desafíos que incluyen nuevas formas de construir la identidad individual, el desplazamiento laboral por las máquinas, nuevas formas de delitos informáticos, y hasta amenazas a la seguridad y a la vida misma de los individuos y de los ecosistemas, entre otros desafíos.

Si podemos afirmar que la computación no es ajena al devenir de la humanidad, como ingenieros y como ciudadanos nos corresponde asumir roles críticos frente los impactos potenciales, positivos y negativos que va teniendo cada nueva tecnología. Para ello, se requiere asumir una formación a partir de la cual los ingenieros de sistemas estén en capacidad dar una mirada crítica a los fenómenos sociales, culturales, políticos y económicos de los entornos en los cuales se inserta cada tecnología. Este es, entonces, un tema que compete a la formación de ingenieros de sistemas y que, se estima, será una de las áreas de mayor expansión, tecnológica y económica, en las próximas décadas: por la creación de nuevas tecnologías cada vez más capaces de imitar tareas intelectuales humanas y por la necesidad de abordar críticamente el impacto de estas tecnologías, como si se tratara de una forma contemporánea de alfabetización, de la que todo ciudadano debería estar al tanto.

---

* Trabajo de Grado

** Facultad de Ingenierías Fisicomecánicas. Escuela de Ingeniería de Sistemas e Informática. Ingeniería de Sistemas. Director: Sonia Cristina Gamboa Sarmiento. Doctora en Educación.

Abstract

Title: Research Seminar on Ethics and Artificial Intelligence: Ethical Implications of the Impact of Artificial Intelligence on Contemporary Society. *

Authors: Juan Pablo Arias Remolina
Yonathan Camilo Benítez Mancipe
Andrés Felipe Cárdenas Zárate
Sonia Marcela Granados Moreno
Neyder Fabian Mosquera Niño**

Key Words: Artificial Intelligence, ethics in artificial intelligence, artificial general intelligence, technological responsibility, artificial intelligence regulation.

Description:

Today, computing focuses primarily on developing efficient hardware and software capable of performing increasingly complex tasks—ranging from the imitation and surpassing of typically human abilities to processes that transcend biological limitations and require technological assistance. While the benefits of computing in all areas of life, industry, and science are immeasurable, the rise of artificial intelligence presents humanity with a range of dilemmas, risks, and challenges. These include new ways of constructing individual identity, labor displacement by machines, novel forms of cybercrime, and even threats to individual and ecological security and survival.

If we can affirm that computing is inseparable from the course of human development, then, as engineers and citizens, we must adopt a critical stance regarding the potential positive and negative impacts of each emerging technology. This demands a type of education through which systems engineers are equipped to critically analyze the social, cultural, political, and economic phenomena that surround technological implementation. This is therefore a matter relevant to the education of systems engineers and is expected to be one of the fastest-growing technological and economic fields in the coming decades—not only due to the creation of increasingly sophisticated technologies capable of imitating intellectual human tasks, but also due to the urgent need to critically address their impact, as a contemporary form of literacy that every citizen should be aware of.

** Faculty of Physical and Mechanical Engineering. School of Systems and Informatics. Systems Engineering Program. Director: Sonia Cristina Gamboa Sarmiento. Ph.D in Education

Introducción

En las últimas décadas, la inteligencia artificial (IA) ha pasado de ser un sueño de los científicos a convertirse en una tecnología que está en todas partes, cambiando rápidamente muchos aspectos de nuestra vida cotidiana. Desde los sistemas que nos sugieren contenido en redes sociales hasta los algoritmos que deciden sobre préstamos, diagnósticos médicos o incluso procesos judiciales. Sin embargo, junto con su capacidad de optimización y eficiencia, emergen también riesgos éticos, sociales y políticos que exigen una reflexión crítica y multidisciplinaria.

Ante las tremendas amenazas diarias que estas tecnologías suponen, es de vital importancia abrir espacios para la reflexión críticas que estén al margen no solo de su lado técnico, sino también de las dimensiones sociales, culturales, económicas y políticas existenciales en que está inserta la IA es la creciente automatización del trabajo, junto con la toma de decisiones mediante algoritmos y el análisis exhaustivo de grandes volúmenes de información, son temas que demandan una reflexión ética inmediata. No basta con desarrollar tecnología que funcione bien; es crucial comprender cómo afecta al mundo y fijar normas que defiendan los derechos humanos, impulsen la igualdad en la sociedad y garanticen un futuro sostenible para el medio ambiente.

Con este seminario, se buscó examinar los impactos sociales, los riesgos inherentes y las formas emergentes de regulación de la IA, a la luz de principios éticos contemporáneos. Estudiando a fondo los fundamentos teóricos y prácticos de la IA, aplicaciones actuales por detrás de la IA, alcances en la gobernanza algorítmica, empleo, derechos humanos y sostenibilidad, se produce una comprensión profunda y crítica del fenómeno y de igual modo se abordaron los debates actuales en torno a las regulaciones nacionales e internacionales, e invita a reflexionar

acerca del lugar de los profesionales de la ingeniería en el diseño, desarrollo y supervisión de sistemas inteligentes destinados a mejorar el bienestar social y los valores democráticos.

1. Generalidades del seminario de investigación

1.1 ¿Qué es el seminario de investigación?

El Reglamento estudiantil de pregrado[5] –REP– propone, entre otras, la modalidad de trabajo de grado denominada seminario de investigación[6], la cual define como "un proceso reflexivo, sistemático y crítico que tiene como propósito fortalecer en el estudiante las habilidades requeridas en el manejo de la información y la comunicación para desarrollar investigación científica" (REP; p. 61). Estas habilidades hacen referencia a la búsqueda y selección de fuentes bibliográficas, a la lectura y escritura crítica de textos y a la discusión ordenada y argumentada, sobre algún tema específico que sea de interés para el campo de conocimiento en el cual están optando por su título profesional.

1.2 Objetivo del Seminario de Investigación

Si bien, la investigación científica implica el desarrollo de otras habilidades (como identificación y formulación de problemas, diseño metodológico, recolección y análisis de datos, entre otras) y la formación gradual de cierta disciplina, el seminario alemán, como se conoce en su origen busca transformar los espacios de docencia tradicionales, en espacios en los que sea posible, al mismo tiempo, aprender sobre determinado tema, explorado al mismo nivel jerárquico

[5] UIS, REP. Acuerdo del Consejo Superior No. 72 de 1982.
[6] Ibid., Capítulo IX, Modificado por el Acuerdo del Consejo Superior No. 004 de febrero 12 de 2007, por el cual se establecen nuevas modalidades y reglamentación para la realización del Trabajo de Grado.

entre profesores y estudiantes, y desarrollar nuevo conocimiento, a la luz tanto de teorías como del estado del arte en el tema, de manera que se logra un cierto nivel de formación científica en los estudiantes, al tiempo que se aborda "el estudio de nuevos objetos de investigación de interés para la Escuela" (REP; Id.).

1.2 Características del Seminario de Investigación

El REP establece que esta modalidad la podrán realizar entre 3 y 5 estudiantes para el mismo seminario, quienes conjuntamente, y bajo la dirección de un profesor, elaboran un plan del Seminario, sobre un tema específico de interés para la Escuela, el cual se organiza por subtemas a estudiar "alrededor del problema seleccionado, la bibliografía a consultar, la programación de sesiones, la asignación de responsabilidades en cada sesión y los relatores respectivos" (REP; p. 68), y "mediante una dinámica que comprende actividades de relatoría, correlatoría, discusión y elaboración de un documento síntesis" (REP; p. 61).

1.2.1 Descripción de los roles

El relator es quien tiene la responsabilidad principal durante una sesión; su función es estudiar a profundidad el tema asignado, preparar una exposición clara y fundamentada, y presentar el contenido al grupo. Este rol exige una preparación rigurosa, ya que el relator debe dominar la bibliografía y los aspectos teóricos y metodológicos del tema para guiar el análisis colectivo.

El correlator complementa la labor del relator, aportando observaciones, críticas constructivas y ampliaciones al tema expuesto. Su función es enriquecer la discusión con aportes

adicionales y ayudar a clarificar puntos que requieran mayor profundidad, facilitando así un debate más completo y crítico.

La discusión es el espacio donde todos los participantes intervienen activamente, aportando sus ideas, cuestionamientos y reflexiones. Este momento es clave para fomentar el pensamiento crítico, la argumentación y la confrontación respetuosa de diferentes puntos de vista, lo que enriquece el aprendizaje colectivo.

Finalmente, el protocolante o responsable del protocolo tiene la tarea de documentar detalladamente cada sesión. Esto incluye registrar el tema tratado, los asistentes, las funciones cumplidas, el desarrollo de la discusión, los acuerdos alcanzados y los interrogantes surgidos. Este documento es fundamental para el seguimiento del seminario y para la elaboración del producto final.

2. Planteamiento del problema

A la luz de los beneficios y los riesgos que implicó el auge de la IA, se pretendía lograr una aproximación para clasificar y caracterizar tanto las formas esperadas de la IA, en términos tecnológicos como de inserción en los contextos sociales, así como los impactos positivos y negativos que podía esperarse de ella en las siguientes décadas y las regulaciones correspondientes.

3. Justificación

En la actualidad, la computación, cuyo interés se centra, principalmente, en desarrollar hardware y software eficiente que logre ejecutar tareas cada vez más complejas, y que van desde la imitación y superación de tareas típicamente humanas hasta procesos que las limitaciones biológicas humanas no les permiten realizarlos sin ayuda tecnológica. Si bien, los beneficios en todas las áreas de la vida, de la industria, de la ciencia, derivados de la computación son incalculables, hoy en día, con el auge de la IA, la humanidad se enfrenta a dilemas, riesgos, desafíos que incluyen nuevas formas de construir la identidad individual, el desplazamiento laboral por las máquinas, nuevas formas de delitos informáticos, y hasta amenazas a la seguridad y a la vida misma de los individuos y de los ecosistemas, entre otros desafíos.

Si podemos afirmar que la computación no es ajena al devenir de la humanidad, como ingenieros y como ciudadanos nos corresponde asumir roles críticos frente los impactos potenciales, positivos y negativos que va teniendo cada nueva tecnología. Para ello, se requiere asumir una formación a partir de la cual los ingenieros de sistemas estén en capacidad dar una mirada crítica a los fenómenos sociales, culturales, políticos y económicos de los entornos en los cuales se inserta cada tecnología.

Este es, entonces, un tema que compete a la formación de ingenieros de sistemas y que, se estima, será una de las áreas de mayor expansión, tecnológica y económica, en las próximas décadas: por la creación de nuevas tecnologías cada vez más capaces de imitar tareas intelectuales humanas y por la necesidad de abordar críticamente el impacto de estas tecnologías, como si se tratara de una forma contemporánea de alfabetización, de la que todo ciudadano debería estar al tanto. Le concierne, entonces, a los ingenieros de sistemas capacitarse no solamente en destrezas

para crear nuevas tecnologías innovadoras, sino en conocimientos y herramientas para crearlas bajo parámetros estrictamente éticas, evaluar críticamente los posibles impactos de tales tecnologías y, por qué no, participar en una necesaria puesta al descubierto y regulación de los impactos que representen riesgos y amenazas para la humanidad o para grupos de esta.

Puntualmente, este seminario propone considerar la IA tanto en sus capacidades tecnológicas actuales e inminentes, así como en los posibles impactos que ya se prevén social y políticamente, y a los principales acuerdos que organizaciones gubernamentales y creadores de tecnología están construyendo.

4. Objetivos del seminario

4.1 Objetivo general

Analizar los impactos, riesgos y regulaciones de la IA a la luz de postulados éticos en concordancia con el estado actual y la proyección de las capacidades tecnológicas de la IA.

4.2 Objetivos específicos

Formular un marco de referencia de la IA que establezca: historia, vertientes disciplinares, capacidades actuales y previstas de la IA.

Formular un marco de referencia de la ética aplicada a la IA, que tenga en cuenta los principales desafíos y sus respectivas regulaciones en desarrollo.

Sintetizar un conjunto de posturas críticas informadas que resulten de las discusiones grupales.

5. Metodología

La modalidad del seminario se acogió a la concepción del seminario alemán[7], que consiste en un enfoque de enseñanza orientado a la formación en investigación, puesto que ubica al estudiante en un rol activo en el cual debía estudiar marcos teóricos y ponerlos en relación con el estado del arte y el contexto actual del asunto en cuestión. Esta puesta en contexto se denomina relatoría y consiste en un documento cuya lectura da lugar a una discusión ordenada, informada y argumentada por parte de todos los miembros del seminario. A partir de la discusión, uno de los miembros del seminario elabora otro documento, denominado protocolo, que registra una síntesis de la discusión, los acuerdos y los diferentes puntos de vista que se manifestaron durante la sesión. Esta modalidad exigió no solamente el interés y motivación de todos los estudiantes, sino también la preparación de cada uno en los temas de cada sesión, de manera que cada nueva discusión retomó lo aprendido en la anterior, y al finalizar el seminario fue posible contar con un compendio equivalente a la revisión teórica y estado del arte de la cuestión.

Para este seminario, una vez acordado el conjunto de temáticas a abordar durante el semestre académico, se desarrolló semanalmente una sesión en la que los miembros asignados previamente presentaron una relatoría que expuso los conceptos correspondientes, en un relato crítico que consideró tanto definiciones y teorías como el estado del arte del asunto. Cada temática estuvo a cargo de un estudiante. Adicionalmente, los estudiantes tuvieron a su cargo la producción de un informe final, compuesto por entre tres y cinco secciones tipo artículo, sobre las temáticas del seminario. Estos informes incluyeron revisión de conceptos, estado del arte, reflexiones

[7] Se desarrolló en la Universidad de Berlín en el siglo XIX y fue propuesto por Wilhelm von Humboldt. Esta modelo buscó darle a la educación superior un enfoque en la investigación libre, la autonomía académica y la integración del aprendizaje y la enseñanza.

argumentativas y la bibliografía completa referenciada en el texto. Los textos se socializaron en las sesiones del seminario, en las cuales se sugirieron ajustes finales para su consolidación.

5.1 Contenido: temas y subtemas

Para el desarrollo del seminario se seleccionaron los siguientes temas y subtemas, con el propósito de abordar de manera integral los aspectos fundamentales relacionados con la ética e IA. La estructura temática busca ofrecer un recorrido progresivo desde los fundamentos y vertientes de la IA hasta sus implicaciones éticas, sociales y regulatorias, facilitando así un análisis crítico y riguroso.

1. Fundamentos y vertientes de la IA.
    - Orígenes y evolución de la IA (desde sus inicios hasta la IA moderna).
    - Principales enfoques de la IA: simbólica, conexionista y generativa.
    - Avances recientes y tendencias emergentes en IA.
2. Capacidades tecnológicas y aplicaciones actuales de la IA.
    - Capacidades actuales de los sistemas de IA: procesamiento de lenguaje natural, visión por computador y aprendizaje profundo.
    - Ejemplos de herramientas y servicios de IA populares, y cuestiones como la desinformación, el plagio y la creación de contenidos falsos.
    - Discusión sobre el potencial y las limitaciones actuales de estas tecnologías.
3. Sesgos, riesgos y desafíos de la IA.
    - Análisis de los sesgos algorítmicos y cómo estos afectan a diferentes grupos sociales.
    - Desafíos éticos: discriminación, privacidad, autonomía y seguridad.

- Ejemplos prácticos de sesgos en IA y estrategias para mitigarlos.
- Clasificación de riesgos: éticos, técnicos, sociales y económicos.

4. Fundamentos de ética aplicada a la IA

- Definición y conceptos básicos de ética.
- Postulados generales de la ética y su relevancia en el contexto de la IA (ética deontológica, utilitarismo, ética del cuidado).
- Principios éticos aplicados a la tecnología: beneficencia, no maleficencia, justicia y autonomía.

5. Ética y aplicación en la IA.

- Consideraciones éticas especificas en el desarrollo y uso de la IA: transparencia, responsabilidad algorítmica, explicabilidad e inteligibilidad.
- Discusión sobre las limitaciones biológicas y el impacto del “efecto perverso” (cuando los resultados no intencionados de la IA contradicen los objetivos éticos)
- Análisis de cómo los valores sociales y culturales influyen en la implementación ética de la IA.

6. Regulación y políticas mundiales sobre IA.

- Revisión de marcos regulatorios internacionales: OCDE, Unión Europea, Estados Unidos, Asia.
- Debates sobre autorregulación vs regulación gubernamental.
- Propuestas de recomendaciones para una regulación ética y efectiva de la IA.
- Discusión sobre los desafíos en la implementación de políticas éticas a nivel global.

7. Ética en el diseño y desarrollo de sistemas de IA *(Ethics by design)*
    - Incorporación de principios éticos desde la etapa de diseño y desarrollo de los sistemas de IA
8. IA en la toma de decisiones y gobernanza algorítmica
    - Influencia de la IA en la toma de decisiones automatizada en sectores clave como salud, justicia, finanzas y recursos humanos.
9. Impacto de la IA en el empleo y la economía
    - Cómo la IA está transformando el mercado laboral, los desafíos del desplazamiento de empleos y las nuevas oportunidades que surgen con la automatización.
    - Cómo la IA puede amplificar las desigualdades existentes, especialmente entre diferentes regiones, clases sociales y grupos minoritarios.
10. IA, derechos humanos y sostenibilidad
    - Impacto de la IA en los derechos humanos, como el derecho a la privacidad, la libertad de expresión y el derecho a la igualdad.
    - Impacto ambiental de los sistemas de IA, considerando el consumo energético de los modelos de IA y los centros de datos.

5.2 Organización de las sesiones

Para el presente caso, se estableció un total de 16 sesiones, estructuradas conforme a los ejes temáticos definidos para el Seminario. Sin embargo, el número de sesiones varió en función del avance y la profundidad con que se abordó cada uno de los capítulos, permitiendo que algunas temáticas requirieran más o menos encuentros.

De igual forma, se garantizó el cumplimiento de la asignación correspondiente a cada uno de los roles estipulados en el REP.

Tabla 1.

Asignación de los roles y organización de las sesiones

Sesión	Tema	Relator	Protocolante
1	Introducción al seminario	Prof. Sonia C. Gamboa	Andrés Felipe Cárdenas
2	Fundamentos y vertientes de la IA	Yonathan Camilo Benítez	Sonia Marcela Granados
3	Capacidades tecnológicas y aplicaciones actuales de la IA	Neyder Fabian Mosquera	Prof. Sonia C. Gamboa
4	Sesgos, riesgos y desafíos de la IA	Andrés Felipe Cárdenas	Yonathan Camilo Benítez
5	Fundamentos de ética aplicada a la IA	Juan Pablo Arias	Neyder Fabian Mosquera
6	Ética y aplicación en la IA	Sonia Marcela Granados	Andrés Felipe Cárdenas
7	Regulación y políticas mundiales sobre IA	Yonathan Camilo Benítez	Juan Pablo Arias
8	Ética en el diseño y desarrollo de sistemas de IA	Neyder Fabian Mosquera	Sonia Marcela Granados
9	IA en la toma de decisiones y gobernanza algorítmica	Andrés Felipe Cárdenas	Prof. Sonia C. Gamboa
10	Impacto de la IA en el empleo y la economía	Juan Pablo Arias	Yonathan Camilo Benítez
11	IA, derechos humanos y sostenibilidad	Sonia Marcela Granados	Neyder Fabian Mosquera
12	Taller para construcción y socialización del informe final	Prof. Sonia C. Gamboa	
13	Construcción y socialización del informe final	Andrés Felipe Cárdenas Juan Pablo Arias	
14	Construcción y socialización del informe final	Sonia Marcela Granados Yonathan Camilo Benítez	
15	Construcción y socialización del informe final	Neyder Fabian Mosquera	
16	Construcción y socialización del informe final	Andrés Felipe Cárdenas Juan Pablo Arias Sonia Marcela Granados Yonathan Camilo Benítez Neyder Fabian Mosquera	

5.3 Matriz de los Protocolos

Las relatorías y los protocolos que documentan el desarrollo de las sesiones del seminario se elaboraron siguiendo los lineamientos establecidos por la Vicerrectoría Académica de la Universidad Industrial de Santander (UIS) desde el año 2007. Estos lineamientos han sido complementados con procedimientos específicos adaptados para la modalidad de seminario de investigación. La matriz propuesta para el protocolo del seminario, que se presenta a continuación, responde a estos criterios y tiene como propósito asegurar un registro ordenado y detallado de cada encuentro.

Tabla 2.

Estructura del protocolo de las relatorías del seminario

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
INGENIERÍA DE SISTEMAS E INFORMÁTICA

SEMINARIO DE INVESTIGACIÓN
ÉTICA E INTELIGENCIA ARTIFICIAL

Sesión: #
Tema: _____
Fecha: _____

1. Asistencia: funciones y responsables de la sesión
   Director: Sonia Cristina Gamboa
   Relator:
   Protocolante:
2. Objetivos de la relatoría:
3. Fuentes de información:
4. Desarrollo del tema (Síntesis de la relatoría)
   - 4.1 Introducción al capítulo o contenido trabajado
   - 4.2 Temas centrales abordados
   - 4.3 Conclusiones Preliminares
5. Desarrollo de la discusión grupal
   - 5.1 Contribuciones relevantes de los participantes
   - 5.2 Preguntas formuladas
   - 5.3 Diferentes contrapuntos o interpretaciones
   - 5.4 Comentarios de la docente

Seminario Ética e Inteligencia Artificial

Figura 1.

Portada del libro síntesis. Seminario Ética e Inteligencia Artificial.

6. Fundamentos y vertientes de la IA

6.1 Fundamentos e ideas clave consideradas para la creación de la IA

A continuación, se ofrece un breve recuento de los referentes disciplinares que aportaron ideas, puntos de vista y técnicas a la IA, para que esta contribuyera, como tecnología, a la resolución de problemáticas de la época, desde su formulación hasta la fecha.

6.1.1 Filosofía y pensamiento lógico

Desde la antigüedad, la filosofía ha sentado las bases para el desarrollo del conocimiento, en general, de la comprensión de las formas de ser del hombre y de la naturaleza, y de la relación entre ellos. En consecuencia, la filosofía ofrece también las bases para la comprensión y el desarrollo de la IA, abordando preguntas fundamentales sobre la mente, el conocimiento y las acciones que se originan en ellos.

Aristóteles introdujo el concepto de silogismo, un método lógico que permite obtener conclusiones a partir de premisas establecidas (cf. Russell & Norvig, 2010, p.5); por ejemplo, si ningún pez puede sobrevivir sin agua y los tiburones son peces, entonces los tiburones no pueden sobrevivir sin agua. Este concepto se constituye en el fundamento de la comprensión y el aprendizaje del pensamiento lógico y de la argumentación y, posteriormente, posibilita la automatización de estos procesos mentales (Gamboa, 2020) y la construcción de sistemas expertos.

Stevens (1984) señala:

> Los sistemas expertos son máquinas que piensan y razonan como lo haría un experto en un dominio específico. Por ejemplo, un sistema experto en diagnóstico médico solicitaría como entrada los síntomas, el estilo de vida, los resultados de pruebas y otros datos relevantes del paciente. Utilizándolos como indicadores, buscaría en su base de datos información que pudiera conducir a la identificación de la enfermedad (p. 40, traducción propia).

En cuanto a la naturaleza de la mente y la inteligencia, es posible considerar dos posturas filosóficas que resultan de relevancia como fundamento para la IA:

El Dualismo de Descartes sostenía que "existe una parte de la mente humana (o alma o espíritu) que está fuera de la naturaleza, exenta de las leyes físicas. Los animales, por otra parte, no poseen esta cualidad dual; podrían ser tratados como máquinas" (citado en Russell & Norvig, 2010, p. 6, traducción propia). Es decir, Descartes manifestaba que los animales no tienen alma ni conciencia, por lo que su comportamiento se debe únicamente a mecanismos físicos, como si fueran máquinas biológicas que responden a estímulos, sin pensar ni razonar. Por ejemplo, un perro ladra porque ha sido condicionado a hacerlo a través de su experiencia, no porque piensa en ello. Por el contrario, un individuo humano puede preguntarse acerca de la corrección de gritar o no en una circunstancia. En esta interpretación, las máquinas pueden simular la inteligencia humana, pero no necesariamente tendrán una conciencia propia, dado que esa conciencia está más allá de los procesos físicos. En la contemporaneidad hay un pluralismo de asistentes virtuales, como *Siri* o *ChatGPT*, que son capaces de responder preguntas y mantener una conversación; más no son conscientes de sí mismos, sino que dan respuesta a partir de patrones de datos.

Una alternativa al dualismo es el materialismo de Hobbes, que "sostiene que el funcionamiento del cerebro según las leyes de la física constituye la mente. El libre albedrío es simplemente la forma en que la entidad que elige percibe las opciones disponibles" (citado en Russell & Norvig, 2010, p. 6, traducción propia). En otras palabras: al pensar y sentir, nuestras neuronas están procesando información en la medida en que las reglas de la física lo permiten (movimiento de partículas, impulsos eléctricos, reacciones químicas, etc.) Hobbes niega la existencia del libre albedrío. Las elecciones no son libres, porque son decisiones que surgen de nuestro cerebro y también de las condiciones del entorno. Lo que creemos que es un proceso de

elección se nos presenta como la entrega de las posibilidades en función de los estímulos y experiencias más inmediatas. Esta idea ha dado paso a la creación de modelos conexionistas y de redes neuronales.

6.1.2 Cibernética y control automático

En el año 1948, Norbert Wiener publicó el concepto de cibernética y la aplicó al estudio de cómo los sistemas biológicos y mecánicos pueden autorregularse a través de la retroalimentación. Su libro *Cybernetics* impulsó la inquietud por máquinas inteligentes. Wiener y sus colaboradores identificaron el comportamiento intencional como un mecanismo regulador para minimizar el error entre el estado de una máquina y su estado objetivo (cf. Russell & Norvig, 2010, p. 15). Por ejemplo, un termostato analiza la temperatura ambiente y la compara con un valor deseado; regula su operación para mantener una temperatura cálida o fría que sea estable. El desarrollo de máquinas inteligentes depende de la autorregulación y del aprendizaje de la experiencia.

6.1.3 La prueba de Turing

En 1950, Alan Turing sugirió que, en lugar de preguntarnos si las máquinas pueden pensar, deberíamos preguntarnos si pueden superar una prueba de inteligencia conductual, conocida como la Prueba de Turing. La prueba consiste en que un programa mantenga una conversación (mediante mensajes escritos en línea) con un interrogador durante cinco minutos. El interrogador debe adivinar si la conversación es con un programa o con una persona; el programa supera la prueba sí engaña al interrogador el 30 % de las veces (Russell & Norvig, 2010, p. 1021, traducción propia).

Según Russell y Norvig (2010):

Turing planteó que, para lograrlo, la computadora necesitaría poseer las siguientes capacidades:

- Procesamiento del lenguaje natural para permitirle comunicarse con éxito en inglés.
- Representación del conocimiento para almacenar lo que sabe o escucha.
- Razonamiento automatizado para usar la información almacenada para responder preguntas y sacar nuevas conclusiones.
- Aprendizaje automático para adaptarse a nuevas circunstancias y detectar y extrapolar patrones.

Además, propuso una versión más avanzada, la Prueba de Turing total, que incluye una señal de vídeo para que el interrogador pueda poner a prueba las capacidades perceptivas del sujeto, así como la oportunidad de que el interrogador pase objetos físicos "a través de la escotilla". Para pasar la Prueba de Turing total, el ordenador necesitará:

- Visión artificial para percibir objetos.
- Robótica para manipular objetos y moverse.

(p. 2, traducción propia).

Turing conjeturó que, para el año 2000, una computadora con una memoria de 109 unidades podría programarse lo suficientemente bien como para superar la prueba. Se equivocó: los programas aún no han logrado engañar a un juez sofisticado. Por otro lado, muchas personas han sido engañadas sin saber que podrían estar chateando con una computadora (Russell & Norvig, 2010, p. 1021, traducción propia).

En 1966, el informático Joseph Weizenbaum, un profesor de informática del Instituto Tecnológico de Massachusetts creó a ELIZA, el primer *chatbot* conversacional, con el objetivo de

demostrar la superficialidad de la comunicación entre humanos y máquinas. Inspirado en la terapia centrada en el cliente de Carl Rogers, ELIZA simulaba ser una psicoterapeuta que respondía mediante la identificación de palabras clave y la reformulación de frases, dando la impresión de comprensión y empatía. A pesar de su simple funcionamiento y su ausencia de memoria o aprendizaje, muchos usuarios creyeron que ELIZA era capaz de comprender sus problemas y hasta cambiaron confidencias con el programa. Weizenbaum quedó sorprendido al ver que esto sucedió y más adelante advirtió de los peligros de dar a las máquinas un determinado nivel de inteligencia y de sensibilidad que no tienen, y alertó de que las decisiones relevantes no deben ser delegadas en ordenadores que no tengan compasión ni un juicio humano (cf. BBC News Mundo, 2018).

Con el paso del tiempo, han surgido muchos sistemas conversacionales, no obstante, se sigue discutiendo si una máquina que pase la prueba de Turing de verdad piensa o no. Muchos filósofos consideran que el hecho de que se supere la prueba no significa para nada que haya un pensamiento real.

Jefferson (1949), citado por Turing y recogido en Russell y Norvig (2010), expresó:

> Solo cuando una máquina pudiera escribir un soneto o componer un concierto gracias a los pensamientos y emociones que sentía, y no por la caída casual de símbolos, podríamos aceptar que máquina es igual a cerebro; es decir, no solo escribirlo, sino saber que lo ha escrito (p. 1026, traducción propia).

Hoy en día los *chatbots* funcionan de una manera mucho más avanzada que antes, pero siguen apoyados en patrones de conversación indicados o predefinidos y en el aprendizaje de datos previos. Si bien pueden dar respuestas coherentes incluso pueden llegar a parecernos «inteligentes» no necesariamente significa que entiendan realmente lo que están diciendo. En este sentido se abre una pregunta, ¿en qué momento la IA puede llegar a pensar como una persona o si únicamente se están desarrollando programas que, aunque responder correctamente a una pregunta sólo parecen

ser inteligentes? Quizás con el paso del tiempo se logre desarrollar máquinas más inteligentes, pero en este sentido de momento sólo hay una buena simulación de la comunicación humana.

El debate de si las máquinas pueden simular el pensamiento ha dado lugar a diferentes formas de entender la IA. A lo largo de su evolución la IA ha seguido diferentes estrategias de modelado de la inteligencia, o bien utilizando la manipulación de símbolos y reglas lógicas o bien emulando el funcionamiento del propio cerebro humano a través de redes neuronales o mediante la generación de ideas nuevas sobre modelos probabilísticos y aprendizaje profundo.

De los anteriores usos de la IA han ido surgiendo las principales corrientes en los que se ha desarrollado la IA, a continuación, se describen en profundidad las principales corrientes y sus características.

6.2 Vertientes de la IA

6.2.1 IA simbólica

La lógica matemática desempeña un papel central para la comprensión y desarrollo de la IA, ya que establece los resultados teóricos mínimos requeridos para la representación del conocimiento y la inferencia lógica. En este sentido, la IA simbólica representa un enfoque que se basa en la manipulación lógica de símbolos que posteriormente pueden ser manipulados mediante las reglas lógicas con el objetivo de inferir nueva información o bien en la toma de decisiones. Este tipo de planteamiento tiene su origen en los trabajos de George Boole, quien ideó la lógica proposicional o lógica booleana, ya que permite representar el razonamiento lógico, a través de términos computacionales (cf. Russell & Norvig, 2010, p. 8).

En el año 1879, el lógico Gottlob Frege amplió el sistema que usó Aristóteles al introducir una lógica de primer orden, que ya incluía objetos y relaciones, y con ello sentó las bases de la

lógica formal que se utiliza hoy día. Posteriormente Alfred Tarski desarrolló una teoría de referencia que conecta los objetos en un sistema lógico con objetos del mundo real. Estas contribuciones han sido fundamentales para la representación del conocimiento en IA (cf. Russell & Norvig, 2010, p. 8).

Las primeras aplicaciones de la IA se basaban en demostrar que las máquinas sí podían razonar lógicamente y resolver problemas matemáticos con sistemas de deducción lógica y algoritmos formales.

Uno de los primeros programas fue *Logic Theorist*, desarrollado en 1956 por Allen Newell, Herbert A. Simon y Cliff Shaw.

Según Russell y Norvig (2010):

> Simon afirmaba que "Hemos inventado un programa informático capaz de pensar de forma no numérica, y con ello hemos resuelto el venerable problema mente-cuerpo". Poco después del taller, el programa fue capaz de demostrar la mayoría de los teoremas del capítulo 2 de Principia Mathematica de Russell y Whitehead. Se dice que Russell se alegró mucho cuando Simon le mostró que el programa había encontrado una prueba para un teorema que era más corta que la de Principia Mathematica (pp. 17-18, traducción propia).

Poco después, en 1957, Allen Newell, Herbert A. Simon desarrollaron el *General Problem Solver* (GPS), un programa más avanzado que intentaba resolver una variedad de problemas, no solo matemáticos. A diferencia de *Logic Theorist*, este programa fue diseñado desde el principio para imitar los protocolos humanos de resolución de problemas. Dentro de la clase limitada de problemas que podía manejar, resultó que el orden en el que el programa consideraba subobjetivos y posibles acciones era similar al en el que los humanos abordaban los mismos problemas. Por lo tanto, GPS fue probablemente el primer programa en incorporar el enfoque de "pensar humanamente" (Russell & Norvig, 2010, p. 18, traducción propia).

En la actualidad, uno de los algoritmos que utiliza la vertiente simbólica es CLIPS (*C Language Integrated Production System*), una herramienta de software de dominio público diseñada para la creación de sistemas expertos y otros programas que requieren soluciones heurísticas. Desarrollada por la NASA entre 1985 y 1996, CLIPS utiliza la programación lógica, empleando encadenamientos hacia delante, lo que, a partir de hechos iniciales, aplica reglas para deducir nuevos hechos. Una de las aplicaciones más frecuentes es el desarrollo de aplicaciones que simulan el razonamiento humano, como el diagnóstico médico, el asesoramiento legal o el desarrollo de planes de empresa (Arrollo, 2022). CLIPS procede en la línea de la IA simbólica, que trabaja con símbolos y reglas lógicas en el marco de un conocimiento representado y procesado. Al utilizar reglas del tipo "si - entonces" para razonar sobre datos estructurados, CLIPS sería una muestra representativa de dicho enfoque simbólico, dado que representa un conocimiento explícito y da lugar a inferencias lógicas en sistemas de tipo experto.

La IA simbólica fue fundamental para la evolución de la IA, pues gracias a este se desarrollaron las primeras herramientas capaces de lograr la imitación del razonamiento lógico humano. Este enfoque permitió avanzar en la construcción de IA más sofisticados, lo que permitió el avance hacia otras corrientes como por ejemplo la IA conexionista. Con todo y que hoy existen técnicas más avanzadas que la IA simbólica, esta última sigue siendo un punto de referencia clave en el desarrollo de los sistemas inteligentes y también continúa sirviendo de inspiración para muchas tecnologías actuales.

6.2.2 IA conexionista

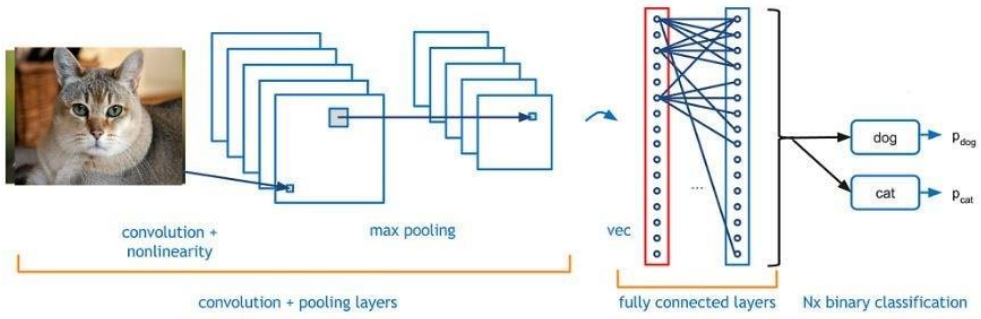
Russell & Norvig (2010) estudian la evolución de las redes neuronales artificiales, las cuales están evidentemente inspiradas en la estructura y el funcionamiento del cerebro humano, el

cual está compuesto por millones de neuronas biológicas interconectadas que se comunican mediante impulsos eléctricos. De la misma manera, las redes neuronales artificiales están formadas por neuronas artificiales que disponen de características visuales y se interconectan entre ellas en función de unos pesos, que son valores numéricos que determinan la intensidad de cada una de esas conexiones (cf. pp. 10-12).

El peso de cada uno de los píxeles es crucial porque determina la influencia de cada una de las características visuales en la clasificación de una imagen. Cuando entrenamos una determinada arquitectura de red neuronal convolucional (CNN) aplicada en la clasificación de imágenes de gatos y perros, la red ajusta los pesos de las conexiones entre las neuronas para aprender qué patrones son más relevantes para distinguir un gato de un perro. Por ejemplo, si determinadas combinaciones de píxeles (p. ej., la forma de las orejas o la textura de la piel) son más representativas de un gato, los pesos de esos píxeles serán más altos que los de los otros. Así, la red neuronal da más importancia a las características esenciales y menos a los detalles irrelevantes, lo cual mejora la precisión (cf. Korlakunta, 2023).

Figura 2.

Proceso de clasificación de imágenes



Nota. Tomado de Korlakunta (2023). Proceso de clasificación de imágenes mediante una red neuronal convolucional para diferenciar entre gatos y perros.

De manera similar, para el caso de procesamiento de texto, las redes neuronales ilustran el escenario en el que se asignan muchos pesos a las palabras, como modo que reside dentro de todos los contextos de la oración. En el caso de los llamados modelos *Transformer*, ponderar permite confirmar qué palabras tienen, en efecto, mayor significado en la frase, aún estén distantes en la secuencia.

Mediante la introducción del *Transformer*, Vaswani et al. (2017) incrementan la traducción automática del inglés al alemán y del inglés al francés. Así, el modelo *Transforme*r permitirá poder procesar textos de forma más eficiente, pudiendo capturar dependencias entre palabras sin importar qué tan distantes estén estas dentro de la secuencia. Dentro del *Transformer*, el mecanismo de *Self-Attention* tiene una importancia fundamental, ya que, permite que el modelo recoja el valor que tienen las distintas palabras dentro de una misma oración, asignando pesos a cada palabra según su valor (cf. pp. 1–4).

La arquitectura de *Transformer* está formada por dos grandes componentes denominados *Encoder* y *Decoder*. En el *Encoder*, todas las palabras de la oración se comparan entre sí a través de *Self- Attention* para determinar con cuáles de ellas se considera que son más relevantes en el contexto. A continuación, en el *Decoder* el modelo utilizará este mecanismo para determinar las palabras en lengua de destino, asegurando que sean coherentes con el texto de origen. Por otro lado, el *Transformer* utiliza *Multi-Head Attention* que consiste en ejecutar múltiples instancias de *Self-Attention* en paralelo y permite que el modelo aprenda diferentes aspectos de significados de una forma más efectiva (cf. Vaswani et al., 2017, pp. 4–9).

En este contexto aparece la IA conexionista, un modelo que realiza el aprendizaje y el procesamiento de la información a través de redes de unidades que se interconectan, como las neuronas.

La base del conexionismo es que el conocimiento no se representa explícitamente mediante símbolos y reglas lógicas como en la IA simbólica, sino por la interacción distribuida de un gran número de unidades simples. Este enfoque pone su origen en 1943, momento en el que Warren McCulloch y Walter Pitts presentaron el primer modelo computacional de una neurona, también conocido como "lógica umbral" en el cual cada neurona artificial tiene vinculadas varias entradas, que pueden ser señales que llegan de otras neuronas o datos de entrada, a las que se les asigna un peso, que es definido manualmente por sus diseñadores. En este caso, la suma de las entradas y el peso se lleva a cabo en una suma ponderada. Si la suma es mayor que el umbral, se produce un efecto activador en esta neurona; en caso contrario, la neurona permanece en un modo inactivo (cf. Russell & Norvig, 2010; pp. 16-17, 731-732). Más tarde, 1958, Frank Rosenblatt desarrolló el Perceptrón, un algoritmo de aprendizaje basado en la lógica umbral, pero incorporando la capacidad de ajustar automáticamente los pesos para poder aprender a clasificar datos y mejorar la clasificación (cf. Russell & Norvig, 2010; pp. 723-733).

Con el paso del tiempo, se pusieron en práctica redes neuronales más complejas, como por ejemplo las redes neuronales multicapa que han demostrado una gran capacidad para poder resolver problemas que van más allá de las redes neuronales de una sola capa.

Las redes neuronales multicapa están formadas por múltiples capas de neuronas artificiales que se organizan en tres tipos de capas fundamentalmente. La primera de ellas es la capa de entrada que se encarga de recibir la información. Más tarde, encontramos la o las capas ocultas que se encargan de procesar y transformar la información, y finalmente, la capa de salida que proporciona la respuesta que finalmente devuelve el sistema.

Cada neurona de estas capas realiza un cálculo relativamente sencillo que está basado en una función de activación, la cual tiene como entrada los valores de las neuronas que considera la

de la capa anterior aplicando pesos y sesgos que son ajustables al mismo. Al principio, estos pesos se asignan de forma aleatoria a las conexiones, pero poco a poco, la red comienza a recibir información desde una sección de entrenamiento ajustando sus conexiones de modo que se acabe minimizando la separación entre las respuestas que proporciona la red y las respuestas que espera debido a las entradas (Russell & Norvig, 2010, pp. 727-733).

Gracias al proceso de ajuste de las conexiones, las redes neuronales multicapa han logrado alcanzar resultados sorprendentes en tareas como por ejemplo el reconocimiento de imágenes, la traducción automática de lenguajes y el desarrollo de asistentes virtuales capaces de responder a preguntas con un alto grado de precisión.

La IA conexionista supuso un cambio considerable en relación a la IA simbólica. Si la simbólica tenía que apoyarse en reglas fijas, la conexionista permitía que los sistemas aprendiesen por sí mismos, lo que los hacía más flexibles y adaptables. Además, este enfoque se convirtió en la raíz de la IA generativa, que es capaz incluso de ir más allá de analizar datos gracias a la posibilidad de generar contenido nuevo (imágenes, textos, música, etc.) y que, sin el avance de la IA conexionista y el aprendizaje automático, muchos de los modelos de IA actuales no serían viables.

6.2.3 IA generativa

Foster (2023) examina cómo, a partir de los datos, no solamente se pueden reproducir, sino que más bien también se pueden innovar y generar nuevos contenidos originales que pueden tener diferentes aplicaciones en ámbitos como el de las imágenes, el texto, la música, etcétera. En este sentido, "el modelado generativo es un área de la IA que consiste en entrenar un modelo para que genere nuevos datos similares a un conjunto de datos dados" (p. 4, traducción propia).

Para ilustrarlo Foster (2023) propone un ejemplo que consiste en tener un conjunto de datos que contenga fotos de caballos. A partir de este conjunto de datos, podemos entrenar un modelo generativo que capture las reglas que sí existen entre las relaciones complejas y a veces abstractas que se pueden establecer entre los píxeles que configuran las imágenes de caballos. Este modelo entrenado puede ser utilizado para obtener muestras de un modelo generativo para generar imágenes compositivas y originales de caballos que nunca habían sido vistas en el conjunto de datos original de caballos (p. 4, traducción propia).

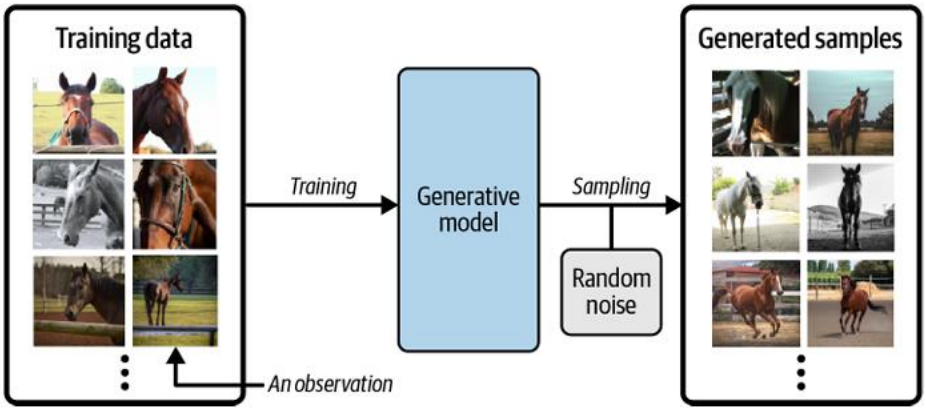
Para construir un modelo generativo, necesitamos un conjunto de datos compuesto por numerosos ejemplos de la entidad que intentamos generar. Estos se conocen como datos de entrenamiento, y uno de estos datos se denomina observación. Cada observación consta de numerosas características. En un problema de generación de imágenes, las características suelen ser los valores individuales de los píxeles; en un problema de generación de texto, las características pueden ser palabras individuales o grupos de letras. Nuestro objetivo es construir un modelo que pueda generar nuevos conjuntos de características que parezcan haber sido creados utilizando las mismas reglas que los datos originales. Conceptualmente, para la generación de imágenes, esta es una tarea increíblemente difícil, considerando la gran cantidad de maneras en que se pueden asignar valores individuales de píxeles y el número relativamente pequeño de tales disposiciones que constituyen una imagen de la entidad que intentamos generar (Foster, 2023, pp. 4-5, traducción propia).

Un modelo generativo también debe ser probabilístico en lugar de determinista, ya que queremos poder muestrear muchas variaciones diferentes del resultado, en lugar de obtener el mismo resultado cada vez. Si nuestro modelo es simplemente un cálculo fijo, como tomar el valor promedio de cada píxel en el conjunto de datos de entrenamiento, no es generativo. Un modelo

generativo debe incluir un componente aleatorio que influya en las muestras individuales generadas por el modelo (Foster, 2023, pp. 4-5, traducción propia).

Figura 3.

Modelo generativo entrenado para generar fotos realistas de caballos.



Nota. Tomado de Foster, 2023, p. 4.

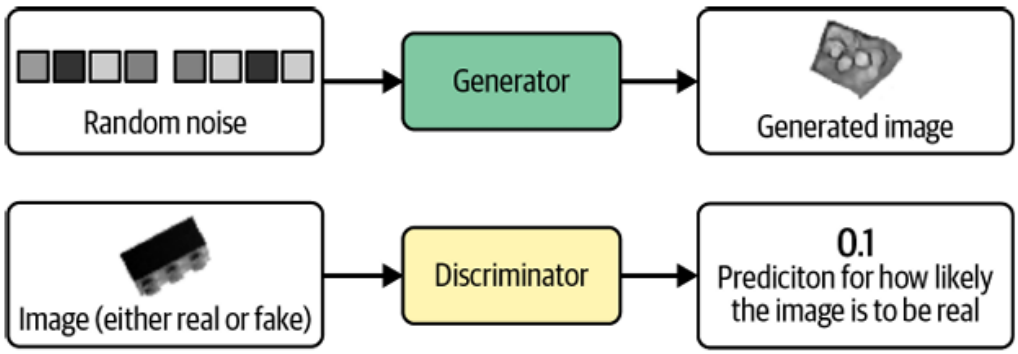
Dentro de la IA generativa podemos encontrar varios tipos de IA generativa. Las más relevantes son las Redes Generativas Antagónicas y el Autocodificador Variacional. Este último hace referencia a lo que Foster (2023) denomina Redes Generativas Antagónicas (GAN): "las cuales son un sistema compuesto por dos redes que deben ganar la competencia entre ambas; un generador y un discriminador. Mientras, el generador intenta generar datos falsos que imiten a los del conjunto de datos original, el discriminador se debe encargar de diferenciar entre los datos que son verdaderos y los que están generados artificialmente." En el mismo sentido, "a medida que ambos modelos mejoran, el generador va llegando más lejos en la adquisición del contenido real que es más parecido a su competencia" (cf. pp. 95-97).

Foster (2023) ilustra el funcionamiento de las redes generativas antagónicas (GANs) a partir de la análoga reflexión entre procesos y resultados de una empresa ficticia que se denomina *Brickkis*, productora de ladrillos, que hace de símil para explicar intuitivamente el funcionamiento

de las GANs. El ladrillero tiene como fin replicar ladrillos de calidad, mientras el jefe del control de calidad se entrena para detectar los ladrillos que no son válidos. A medida que ambos van entrenando sus habilidades en un ciclo de competición, la producción de ladrillos equivales a una calidad que no se puede distinguir de la de los ladrillos originales. De ahí la representación de las capacidades del generador (el ladrillero) y del discriminador (el jefe del control de calidad) en una GAN, por un lado, el generador produce datos sintéticos creíbles y de otro lado, el discriminador entrena para distinguir los datos sintéticos de los reales (cf. pp. 95-97).

Figura 4.

Entradas y salidas de las dos redes en una GAN.



Nota. Tomado de Foster, 2023, p. 97.

Según Foster (2023), la clave en las GAN radica en la forma en que alternamos el entreno de ambas redes, de manera que una vez el generador se vuelve hábil engañando al discriminador este debe adecuarse y así poder continuar siendo capaz de identificar correctamente qué observaciones son falsas; el generador, simultáneamente manteniendo su objetivo de lograr engañarle, comienza a buscar formas para engañarle, y así el ciclo continúa (cf. p. 97).

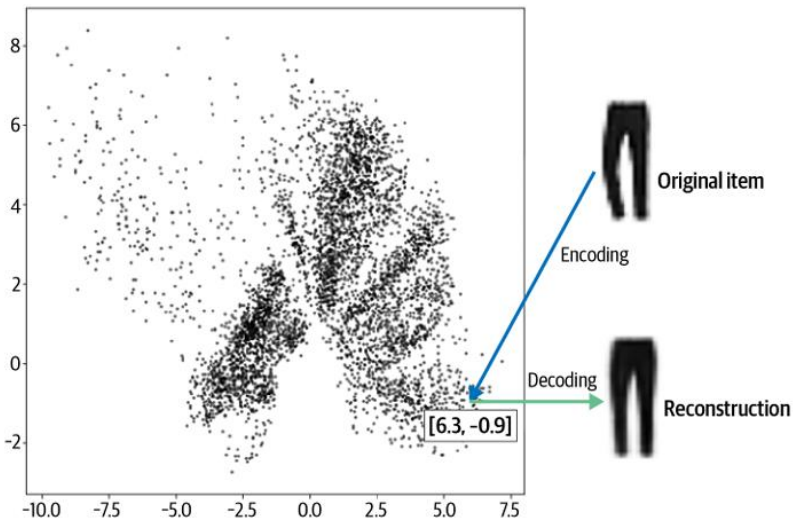
Por otro lado, Foster (2023) en los Autocodificadores Variacionales (VAE), es un modelo generativo que puede aprender a través de una representación latente de los datos de entrada, debido a la compresión de la información que contenga en un espacio de menos dimensiones. Con

esta representación se puede generar nuevos datos, al momento de decodificar puntos aleatorios dentro de ese espacio latente (cf. pp. 59-94).

Para ilustrarlo, Foster (2023) presenta la metáfora de un estilista llamado Brian y un armario infinito para explicar cómo funcionan los autocodificadores variacionales. En esta historia, tú colocas tu ropa en distintas ubicaciones del armario (proceso de codificación), y Brian, a partir de esas coordenadas, es capaz de coser de nuevo cada prenda (proceso de decodificación). Con el tiempo, incluso puedes darle ubicaciones vacías y Brian genera nuevas prendas que nunca han existido. Esto ilustra cómo los autocodificadores variacionales no solo reconstruyen datos, sino que pueden generar muestras nuevas manipulando el espacio latente. A diferencia de los GANs, donde dos redes compiten, los autocodificadores variacionales aprenden a generar variaciones realistas mediante una representación estadística del espacio de características (cf. pp. 59-62).

Figura 5.

Prendas en el armario infinito.



Nota. Cada punto negro representa una prenda. Tomado de Foster, 2023, p. 61.

Cada ubicación en el armario está representada por dos números (es decir, un vector 2D). Por ejemplo, los pantalones de la Figura 3-2 están codificados hasta el punto [6.3, –0.9]. Este

vector también se conoce como incrustación, ya que el codificador intenta incrustar en él la mayor cantidad de información posible para que el decodificador pueda producir una reconstrucción precisa (Foster, 2023, pp. 61-62, traducción propia).

Un autocodificador es simplemente una red neuronal entrenada para codificar y decodificar una prenda, de modo que el resultado de este proceso sea lo más parecido posible a la prenda original. Fundamentalmente, puede usarse como modelo generativo, ya que podemos decodificar cualquier punto del espacio 2D que queramos (en particular, aquellos que no son incrustaciones de prendas originales) para producir una prenda nueva (Foster, 2023, pp. 61-62, traducción propia).

6.3 Limitaciones iniciales de la IA

En los primeros años de la IA, muchos investigadores estaban muy entusiasmados por los avances iniciales. Algunos, como Herbert Simon, creían que en poco tiempo las máquinas podrían pensar, aprender y resolver problemas igual que los humanos. Sin embargo, con el paso del tiempo, estas predicciones no se cumplieron, y eso provocó una gran desilusión (cf. Russell & Norvig, 2022, pp. 39-40).

Entre 1966 y 1973, se hizo evidente que muchos sistemas de IA que funcionaban bien en ejemplos simples no eran útiles para resolver problemas más complejos del mundo real. Esto ocurrió por dos razones principales. Primero, muchos programas se diseñaron basándose solo en ideas generales sobre cómo pensamos los humanos, sin estudiar a fondo lo que realmente se necesita para resolver un problema de forma precisa. Segundo, no se entendía bien lo difícil que eran algunos de estos problemas. Al intentar resolverlos, los sistemas no podían manejar la enorme cantidad de posibles combinaciones y caminos a seguir (cf. Russell & Norvig, 2022, pp. 39-40).

Una crítica importante a la IA en esa época fue el informe Light Hill (1973), que llevó al gobierno del Reino Unido a dejar de apoyar económicamente la investigación en este campo, excepto en dos universidades (cf. Russell & Norvig, 2022, pp. 39-40).

Otro golpe vino de las redes neuronales tempranas, llamadas perceptrones. Aunque en teoría podían aprender, en la práctica eran muy limitadas y no podían hacer cosas tan simples como distinguir si dos entradas eran diferentes. Por esto, durante muchos años, se detuvo casi por completo la investigación en este tipo de sistemas, a pesar de que ya existían métodos prometedores, como el aprendizaje por retro propagación, que más adelante causarían un gran avance en la IA (cf. Russell & Norvig, 2022, pp. 39-40).

6.4 Avances recientes

De acuerdo con Parra (2024), *AlphaFold*, un programa diseñado por *DeepMind*, ha revolucionado la biología molecular al predecir con fiabilidad la estructura tridimensional de las proteínas a partir de su secuencia de aminoácidos. Este programa ha permitido dar el gran salto hacia el estudio de enfermedades provocadas por las proteínas mal plegadas, ya que antes de la aparición de *AlphaFold* el determinar la estructura también requería técnicas muy costosas y lentas como era la cristalografía de rayos X. Como resultado, la búsqueda de nuevos tratamientos médicos se ha visto enormemente beneficiada por *AlphaFold*.

AlphaFold usa redes neuronales en profundidad que han sido entrenadas a partir de bases de datos de proteínas ya caracterizadas. Usa información evolutiva y modelos de aprendizaje profundo para generar representaciones de dichas estructuras; desde un punto de vista experimental, se conseguían resultados que rivalizaban con las técnicas de laboratorio. Este avance ha permitido acelerar el proceso de descubrimiento de fármacos, el estudio de distintas condiciones

producidas por enfermedades genéticas o la ingeniera de nuevas enzimas para aplicaciones industriales y ambientales (Parra, 2024).

Gracias a su influencia en el ámbito de la ciencia, David Baker, Demis Hassabis y John Jumper, los principales investigadores detrás de *AlphaFold* fueron premiados con el Premio Nobel de Química en este mismo 2024. A partir de sus investigaciones se ha inventado un punto de inflexión en la investigación biomédica, probando que la IA puede ayudar a solucionar diferentes problemas de gran complejidad científica y transformar la historia futura de la medicina y de la biotecnología (Parra, 2004).

Por otro lado, también en otros dominios se están llevando a cabo iniciativas, por ejemplo, en la fusión nuclear, donde la IA comienza a marcar una diferencia de importancia. La fusión nuclear se considera por muchos la fuente de energía limpia, segura y sostenible del futuro, pero presenta problemáticas técnicas y científicas plurales. En este sentido, De Vries (2024) explican que los avances en IA contribuyen a la aceleración del desarrollo de tecnologías fundamentales en reactores de fusión. Entre ellos, el control del plasma, la mitigación de disrupciones y la rápida interpretación de datos experimentales. Para acortar el camino hacia su aplicación práctica, el consorcio EURO fusión ha lanzado 15 nuevos proyectos de investigación que integran IA y aprendizaje automático como herramientas estratégicas para resolver estos retos.

Entre los desafíos que se abordan con IA están la reconstrucción de señales de reflectometría, la aceleración de simulaciones de estabilidad del plasma, la detección de inestabilidades, y la clasificación automática de imágenes y patrones físicos en grandes volúmenes de datos. Uno de los propósitos comunes es la aplicación de estos avances en escenarios óptimos, como el control en tiempo real del plasma y la predicción de condiciones óptimas en este tipo de reactores (ejemplo: ITER y DEMO, este último como el futuro reactor de fusión en Europa (De Vries,

2024)); es decir, estos esfuerzos situarían la IA como una herramienta indispensable para acelerar el desarrollo del tipo de fusión nuclear planteado, cambiando así una expectativa improbable por una solución energética en el siglo XXI (De Vries, 2024).

6.5 Tendencias emergentes de la IA

La IA está avanzando rápidamente, pero también presenta desafíos que deben abordarse para evitar riesgos en el futuro. Russell (2019) considera que la falta de comprensión real de la IA es un problema porque las máquinas simplemente optimizan objetivos sin entender su verdadero significado o las consecuencias de sus acciones. El estricto cumplimiento de una especificación de los objetivos puede llevar a una IA a producir resultados poco deseados e incluso potencialmente peligrosos, de forma particular si no se han definido adecuadamente dichos objetivos, o peor aún, si la IA hay de seguirlos con una rigidez tal que ignoren explícitamente valores humanos. Imaginemos por un instante que pedimos a una IA que optimice la circulación del tráfico de una ciudad. Siguiendo un planteamiento de optimización estricto, la IA podría determinar que el objetivo único sea que los coches se muevan lo más rápido posible, para llegar tal extremo que concluyera eliminando todos los semáforos, y además que, en un ejercicio de extrema obsesión por llegar al objetivo, decidiera incluso prohibir el paso de los peatones, desconocedores por completo de que sus decisiones ponen en peligro a las personas que deambulan por las calles. Como la IA carece de sentido común y de juicio moral, de la misma manera que aplicaría un procedimiento de optimización estricto, llegaríamos a este punto, a unas soluciones que, aunque estrictamente hablando cumplen un objetivo de circulación del tráfico hacen que el tráfico de la ciudad sea una pesadilla para sus habitantes y lo humanamente intolerable (cf. pp. 136-140).

Para hacer frente a esto, Russell (2019), opta por una estrategia conocida con el nombre de Aprendizaje por Recompensa Inversa, la IA observaría de qué manera los seres humanos valoran tanto una óptima rapidez como una óptima seguridad, se daría cuenta de que, si bien queremos disminuir el tráfico, también colocamos pasos peatonales y semáforos para resguardar a las personas, es decir, la IA podría llegar a un equilibrio entre eficiencia y seguridad sin necesidad de asegurar instrucciones explícitas para cada posible escenario (cf. pp. 190-196).

Otra tendencia significativa que extrae el autor es la necesidad de una regulación internacional de la IA, pues de no existir normas claras podría emplearse de manera irresponsable, en procesos de manipulación de la información en redes sociales o incluso, en la creación de armas autónomas que elijan a qué atacar u omitir su decisión de ataque sin intervención humana. Es por esta razón que países y empresas tienen que crear normas para que la IA no sea utilizada de forma temeraria (cf. pp. 103-113).

Este recorrido a través de los fundamentos y vertientes de la IA deja claro cómo la IA ha pasado de ser una mera idea filosófica para convertirse en una tecnología presente en muchos aspectos de nuestra vida cotidiana. Desde el pensamiento lógico de Aristóteles hasta los modelos más avanzados, lo que ha llevado a cabo la evolución de la IA ha estado marcada por avances técnicos espectaculares, pero también por retos y desventajas continuas.

Si bien tenemos en la actualidad sistemas que pueden conversar, traducir, generar, crear contenido o incluso predecir estructuras de proteínas, la verdad es que estamos aún muy lejos de cumplir la Prueba de Turing; muchos modelos pueden engañar por un instante, pero no comprenden lo que dicen, ni tienen conciencia, ni tienen emociones como las que puede tener un ser humano.

7. Capacidades tecnológicas y aplicaciones actuales de la IA.

La IA ha sido, tal vez, la tecnología más transformadora del mundo actual, su adopción masiva en diversos sectores de la vida ha reestructurado los modelos de interacción humano-tecnológica y redefinido los procesos de toma de decisiones a escala global. Los avances en áreas como el aprendizaje profundo, el procesamiento de grandes volúmenes de datos y la arquitectura computacional han permitido que los sistemas de IA logren niveles de precisión y eficiencia que antes parecían inalcanzables. En la medida que este tipo de tecnología continúa en evolución, se plantean cada vez mejores oportunidades de resolución de problemas y situaciones que aquejan a la humanidad, a la vez que se presentan retos éticos que es necesario considerar, entre ellos, la necesidad de un desarrollo con enfoque consistente y responsable. En este capítulo se exploran las capacidades tecnológicas actuales que tiene la IA y sus aplicaciones en diferentes sectores.

7.1 Capacidades tecnológicas de la IA

El diseño de sistemas de IA está fundamentado en un principio clave: creación de sistemas orientados a objetivos específicos. Este enfoque implica que cada sistema debe ser diseñado para cumplir una meta determinada, lo cual constituye un propósito central. Para lograrlo, la IA debe integrar capacidades que le permitan modificar sus operaciones internas y adaptarse de forma dinámica a las condiciones cambiantes del entorno. Un sistema de IA comienza con la identificación precisa de los objetivos que se deben alcanzar. Estos objetivos orientan todas las decisiones y acciones del sistema, asegurando que cada operación esté alineada con el propósito definido.

Los sistemas de IA deben ser capaces de analizar datos en tiempo real y ajustar sus procesos internos para responder a los cambios del entorno. Por ejemplo, los agentes basados en objetivos emplean retroalimentación del entorno para optimizar estrategias y mantener su alineación con la meta definida —un agente de IA es un sistema capaz de realizar tareas de forma autónoma; los agentes basados en objetivos están diseñados con el propósito de alcanzar metas específicas mediante la planificación estratégica y la adaptación dinámica, evaluando las consecuencias futuras de sus acciones y ajustando su comportamiento—.

Por otro lado, los sistemas de IA precisan funcionar con un elevado nivel de autonomía, lo cual les permitirá tomar decisiones sin intervención humana directa. Lo cual implica procesar información, aprender de interacciones pasadas y modificar el comportamiento de los sistemas de IA de forma que se busquen maximizar la eficiencia, la cual, junto con la toma de decisiones, también se busca maximizar el logro del objetivo. Al mismo tiempo, el proceso de desarrollo de este tipo de sistemas es de forma ineludible iterativo, es decir, que la creación y la mejora continua no se da en un proceso lineal, sino que se da en ciclos repetitivos de prueba y error. Basándose en estos ciclos de pruebas continuas, de retroalimentación y ajuste, los sistemas mejoran su comportamiento y se adaptan mejor a las necesidades cambiantes del entorno.

Las capacidades actuales de los sistemas de IA engloban las habilidades y funcionalidades en su desarrollo más reciente. Esto incluye el procesamiento avanzado de la información, el aprendizaje a partir de grandes volúmenes de datos, la generación automatizada de contenido y la toma de decisiones en diversas áreas de aplicación.

7.1.1 Procesamiento de Lenguaje Natural — PLN o NLP, por sus siglas en inglés —

Según Daniel Jurafsky y James H. Martin (2020), "el procesamiento de lenguaje natural es un área de estudio que se ocupa de la interacción entre las computadoras y el lenguaje humano, incluyendo la comprensión, el análisis y la generación del lenguaje de modo que una computadora pueda realizar tareas que en la actualidad requieren inteligencia humana". Por otro lado, Manning y Schütze (1999) definen el PLN como "el uso de técnicas computacionales para hacer que el lenguaje humano sea inteligible[8] para las computadoras".

Para que las máquinas comprendan, interpreten y generen lenguaje natural[9], el PLN combina la lingüística computacional —que busca la representación del lenguaje humano en modelos computacionales— con modelos estadísticos —que analizan patrones de frecuencia y distribución de elementos lingüísticos para predecir comportamientos del lenguaje —. A esto se suma el aprendizaje automático —técnicas de IA que permiten a los algoritmos identificar y aprender patrones complejos a partir de grandes volúmenes de datos lingüísticos— y el aprendizaje profundo —subcampo del aprendizaje automático que utiliza redes neuronales artificiales (RNA) para modelar representaciones jerárquicas del lenguaje—.

Esa fusión de varios enfoques dio lugar a la posibilidad de que los sistemas de IA procesaran datos lingüísticos masivos (tanto orales como textuales), que permitieron por ejemplo ejecutar traducción automática manteniendo matices semánticos, es decir, el significado preciso, pero a su vez las connotaciones culturales y las intenciones comunicativas del texto en la traducción, análisis de sentimientos; es decir la detección de la emoción o las distintas emociones en textos o discursos y sistemas inteligentes de asistencia virtual que generan diálogos que son

[8] Algo que puede ser entendido.

[9] Formas de comunicación utilizadas por los seres humanos, como el habla o la escritura, con estructura gramatical, vocabulario y que puede ser usada en diferentes contextos.

pertinentes. Estos sistemas manuales de asistencia virtual logran automatizar tareas, interactuar con un usuario y optimizar procesos en distintos contextos.

La tecnología de PLN ha evolucionado significativamente en los últimos años gracias a los modelos de redes neuronales profundas —término que se abordará a detalle más adelante—. Un ejemplo de éstos son los modelos de lenguaje avanzado: modelos como GPT-4[10] —modelo de lenguaje de IA desarrollado por *OpenAI*, basado en arquitectura de *Transformers*[11] —destaca por su capacidad en comprensión y generación de texto, razonamiento lógico —resuelven problemas matemáticos básicos y siguen instrucciones estructuradas mediante patrones estadísticos aprendidos, no mediante comprensión consciente— y procesamiento de información compleja; se utiliza en una variedad de aplicaciones, desde generación de texto hasta *chatbots*. Así mismo, herramientas como *DeepL*[12] y *Google Translate*[13], utilizan modelos neuronales y PLN para proporcionar traducciones en diferentes idiomas en tiempo real, lo que ha permitido reducir las diferencias idiomáticas al rededor del mundo. *Amazon Alexa*[14], *Apple Siri*[15], *Microsoft Cortana*[16], *IBM Watson Assistant*[17] y *Google Assistant*[18] son claros ejemplos de asistentes virtuales que emplean PLN para responder consultas, automatizar tareas y mejorar la interacción humano-máquina en dispositivos inteligentes, proporcionan una experiencia de usuario más fluida e intuitiva, integrando capacidades de comprensión de lenguaje.

---

[10] OpenAI. (2023). GPT-4. https://openai.com/index/gpt-4/

[11] Arquitectura de red neuronal que transforma o cambia una secuencia de entrada en una secuencia de salida.

[12] DeepL GmbH. (s. f.). DeepL Translator. https://www.deepl.com/translator

[13] Google LLC. (s. f.). Google Translate. https://translate.google.com/

[14] Amazon. (s. f.). Alexa. https://developer.amazon.com/en-US/alexa

[15] Apple Inc. (s. f.). Siri. https://www.apple.com/siri/

[16] Microsoft. (s. f.). Cortana. https://www.microsoft.com/en-us/cortana

[17] IBM. (s. f.). Watson Assistant. https://www.ibm.com/cloud/watson-assistant

[18] Google LLC. (s. f.). Google Assistant. https://assistant.google.com/

7.1.2 Visión por computador

Durante años, la visión ha sido un tema importante para la filosofía y la ciencia. En la antigüedad, Aristóteles (siglo IV a.C), quien abordó la visión desde una perspectiva filosófica y naturalista en su obra *De Anima,* afirmaba que la vista era el sentido más importante, pues permitía percibir los objetos a través de su color, y a su vez, que los colores son causados por la luz y el medio de transmite la imagen. Por otra parte, Platón, en su obra *Timeo*, sostenía que la visión era posible gracias a un "fuego interno" que emanaba desde los ojos, el cual se combinaba con la luz exterior para producir la percepción visual (Platón, Timeo, 45b–46a). A mediados del siglo XI, Alhacén propuso una teoría sobre la óptica que se basaba en la idea de que los objetos emiten rayos de luz que son captados por el ojo, estableciendo aquí las bases de la óptica moderna (Smith, 2001, pp. 58-59).

Ya en la filosofía moderna, Kant afirmaba que la visión no solo era recibir información sensorial, sino que el cerebro estructura la experiencia visual basándose en conceptos previos. En su obra *Crítica de la razón pura (en alemán: Kritik der reinen Vernunft)*, Kant sostiene que el conocimiento humano surge de la síntesis entre la intuición sensible y las categorías del entendimiento, lo cual implica que toda experiencia, incluida la visual, está mediada por estructuras mentales a priori (Kant, 2005, pp. A50–B74). Esto tuvo bastante influencia en la idea de que la visión, aparte de ser un fenómeno físico, también era un proceso interpretativo del cerebro.

Pero ¿de qué manera el cerebro interpreta el mundo? La actualidad apunta a que la visión es un proceso neurofisiológico complejo: el ojo enfoca la luz en la retina; los fotorreceptores (conos y bastones) cambian la luz en señales eléctricas; estas señales llegan hasta la corteza visual en el lóbulo occipital a través del nervio óptico; el cerebro realiza una reconstrucción de la imagen y es capaz de determinar colores, formas y movimiento. Estos principios han sido utilizados como

fundamentos para crear sistemas artificiales que intentan imitar la forma de interpretación de imágenes y de escenas por parte de los seres humanos (Purves et al., 2018).

Para Sucar y Gómez (2000), la visión computacional es “el estudio de los procesos que permite reconocer y localizar objetos en el ambiente mediante el procesamiento de las imágenes” (p. 1). Por otra parte, Szeliski (2011) define la visión por computador como la disciplina que estudia la forma como las computadoras interpretan y comprenden imágenes y videos digitales. Esta disciplina reúne áreas como la informática, la estadística y la óptica, integrando técnicas avanzadas en aprendizaje profundo, reconocimiento de patrones, análisis estadístico y modelos matemáticos que facilitan el procesamiento de la información.

El procesamiento de imágenes, eje central de la visión por computador, empieza con la captura digital de la realidad física. Tal conversión tiene lugar con dispositivos como cámaras CCD, sensores *LiDAR* o sistemas de imagen térmica. De esta forma, la luz ambiental se convierte en representaciones numéricas estructuradas en matrices bidimensionales. Cada píxel (la unidad mínima de información) conserva valores numéricos que representan atributos como la iluminación (que será el caso de la escala de grises), o la composición del color (por ejemplo, los espacios RGB o HSV). Esta representación, sin embargo, está lejana de ser útil para una máquina: una fotografía de un gato, para un algoritmo, no es más que una enorme matriz de números y no tiene, cuando mucho, un significado atribuido. Aquí se encuentra el auténtico reto de la visión artificial: convertir esta sopa numérica en conceptos abstractos y relaciones espaciales (Szeliski, 2022).

Los logros recientes en la creación de hardware (juntamente con las unidades de procesamiento gráfico GPU y los sensores de imagen de alta resolución) y los avances en la creación de algoritmos de aprendizaje automático y redes neuronales profundas, han dado lugar a

la situación actual de la visión por computador que sigue extendiéndose y aplicándose a sectores muy diversos. La visión por computador ha alterado radicalmente las actividades que se desarrollan en la medicina moderna: se usa como un ayudante en el diagnóstico y tratamiento de las patologías. El terreno de las imágenes médicas (radiografías, RM, TC, ultrasonidos, endoscopias) es utilizado como un escenario en el que los algoritmos de IA funcionan como detectives microscópicos que tienen la capacidad de reconocer patrones que escapan incluso a la mirada del radiólogo más experto. Estos sistemas de IA no solo reconocen las células individualmente, sino que reconocen la morfología de estas, su distribución espacial y las características bioquímicas indirectas que se asocian a sutilezas variacionales en la textura y el contraste que tienen las imágenes.

Para dar un ejemplo, en el caso de las biopsias digitalizadas, los sistemas de IA examinan miles de células simultáneamente, detectando núcleos deformes o agrupaciones anormales que podrían indicar cáncer. Plataformas como *Zebra Medical Vision* han desarrollado algoritmos que analizan la arquitectura tisular en biopsias hepáticas, identificando patrones de fibrosis característicos de enfermedades crónicas con una precisión comparable a la de patólogos con décadas de experiencia. Por su parte, *DeepMind* ha revolucionado la oftalmología con sistemas que escanean tomografías de retina para predecir el riesgo de ceguera diabética, analizando minúsculas hemorragias y exudados en la mácula que podrían pasar desapercibidos en un examen rutinario.

Lo fascinante de estas tecnologías es la posibilidad de aprender patrones multiescalar, es decir, estructuras que persiguen la solución de una serie de retos mediante la adaptación a las diferentes escalas de observación. Explotan el potencial de las redes neuronales convolucionales, que actúan a modo de lupas digitales que testan las imágenes mediante una variedad de escalados

en el nivel de zoom, es decir, en el tipo de abstracción; primero detectan bordes y texturas básicas, después montan estas características en formas reconocibles (como, por ejemplo, un vaso sanguíneo o una lesión), y, por último, concentran toda esta información para emitir un diagnóstico en términos de probabilidades. Además, la visión por computadora guía robots quirúrgicos como el sistema quirúrgico Da Vinci que ofrece una visión tridimensional aumentada en la que se distingue con precisión milimétrica los tejidos sanos de aquellos que presentan cáncer. Lo que los sistemas de las tecnologías no persiguen es sustituir al médico, sino magnificar sus capacidades, actuar como asistentes de segunda opinión, bien señalando una zona sospechosa en una mamografía o bien calculando el volumen exacto de un tumor pulmonar en un estudio de tomografía. Su verdadera potencia radica en la detección precoz: algoritmos predictivos que estudian las imágenes de pacientes asintomáticos que muestran signos discretos de enfermedades neurodegenerativas como el Alzheimer a partir de patrones específicos de atrofia del cerebro en resonancias magnéticas.

La visión por computador también ha transformado la industria automotriz, de transporte y de seguridad, impulsando avances que van desde la conducción autónoma hasta sistemas de seguridad biométrica. En el caso de los vehículos autónomos, empresas como *Tesla* y *Waymo* han desarrollado sistemas que integran cámaras de alta resolución, sensores *LiDAR* y radares para crear una percepción tridimensional del entorno. Estos sistemas no solo identifican señales de tráfico y peatones, sino que analizan patrones de movimiento complejos; un auto puede detectar si un ciclista está extendiendo el brazo para girar, o si un peatón muestra intención de cruzar la calle mediante su postura corporal y dirección de mirada. La clave está en las redes neuronales que procesan imágenes en tiempo real, clasificando objetos con técnicas como detección por contornos

—por ejemplo: para reconocer la silueta de camión— y análisis de texturas —para diferencias entre la calle mojada o seca—.

En el campo de la seguridad biométrica, el reconocimiento facial ha evolucionado más allá del simple mapeo de rasgos. Por ejemplo, tecnologías como el *Face ID*[19] de *Apple* utilizan un sistema de proyección de puntos infrarrojos que crea un mapa 3D de la cara del usuario, analizando más de 30,000 puntos invisibles que incluyen la curvatura de las cejas, la profundidad de las órbitas oculares y la topografía única de la nariz. Este modelo matemático se almacena cifrado en el dispositivo, comparando en milisegundos cada nuevo intento de desbloqueo con el patrón original. Además de los rasgos estáticos, algunos sistemas incorporan biometría conductual —tecnología que identifica a individuos mediante el análisis de patrones únicos en su comportamiento al interactuar con dispositivos o entornos —, detectando pequeños gestos como el parpadeo o los movimientos de los labios para evitar suplantaciones con fotografías o máscaras.

7.1.3 Aprendizaje profundo — Deep Learning—

El aprendizaje humano, con independencia de si su definición se toma de un enfoque biológico, psicológico, pedagógico, etc., podría decirse, en términos muy superficiales, que consiste en la adquisición o aprehensión de datos mediante la vivencia de experiencias, a los cuales se les otorga sentido mediante la intuición y la capacidad de comprensión. En una computadora, el aprendizaje consiste en la construcción de una red neuronal que permite un procesamiento adecuado de la información. La red neuronal procesa los daos, pero no de la forma en la que lo haría un ser humano. ¿En qué se diferencia, entonces, el aprendizaje de un humano del aprendizaje de una computadora? En el primero hay una dación de sentido que está determinada por su

[19] Apple Inc. (2025). Use Face ID on your iPhone or iPad Pro. Apple Support. https://support.apple.com/en-us/108411

capacidad para razonar, para comprender y por los juicios, creencias, emociones y vivencias que vienen al caso cuando se lleva a cabo esta síntesis. La computadora, en cambio, solo está diseñada para procesar grandes volúmenes de datos de forma rápida y con una cierta orientación determinada por la estructura matemática de los algoritmos que la procesan.

Para comprender mejor el *Deep Learning* **—** DL **—** se debe tener en cuenta su ubicación dentro del campo de la IA. El DL forma parte de la IA, así como del aprendizaje automático o también llamado *Machine Learning*[20] —ML —. Es decir, el DL es un subconjunto del ML, que a su vez es un subconjunto de la IA. Tiene sus bases en las redes neuronales artificiales profundas **—**DNN**—** diseñadas para imitar el funcionamiento del cerebro humano mediante la combinación de entrada de datos, pesos y sesgos. Para comprender mejor las redes neuronales, es necesario considerar como funciona el cerebro humano: está formado por miles de millones de células llamadas neuronas, cada una de estas neuronas recibe señales de otras neuronas y, con cierto criterio, decide si apagar esa señal o pasar la información a la siguiente neurona; el conjunto de neuronas conforma una red de conexiones —sinapsis— que forman circuitos, que permite recordar, aprender y tomar decisiones.

Las redes neuronales profundas o DNN Deep neural networks son modelos de redes neuronales para resolver tareas y no redes neuronales biológicas; son nodos virtuales que tratan de simular el funcionamiento de una o más neuronas en el cerebro. Tal como ya hemos indicado antes, cada nodo tiene una tarea específica a la que se dedica y están organizados por capas; existe una capa inicial de nodos que recibe la información del exterior (un texto, por ejemplo; datos numéricos, por ejemplo; o bien una imagen, por ejemplo.) A continuación, tenemos las capas

[20] Arthur L. Samuel lo define como el campo de estudio que da a las computadoras la capacidad de aprender sin ser programadas de manera explícita.

intermedias que procesan la información y finalmente una capa última que da la respuesta de la red. Cada nodo de una capa se conecta con los nodos de la siguiente capa de forma que se establece una cadena de decisiones; así, si por la red se quiere identificar si hay un gato en una imagen, la información va pasando de capa y nodo; algunos nodos identifican colores, otros nodos detectan formas y bordes y, por ejemplo, los nodos de la capa de salida darán como respuesta "sí", "no", "no lo sé", etc. Estos modos más profundos de la red intentan reconocer patrones más complejos, emitiendo al final una respuesta afirmando o negando la existencia de un gato en la imagen. Pero ¿cómo aprenden los nodos? La red neuronal realiza su aprendizaje ajustando las conexiones entre nodos, similar a la forma como el cerebro humano fortalece ciertas sinapsis cuando se aprende algo nuevo, si la red comete un error ajusta la fuerza de las conexiones entre nodos para que pueda acercarse más a la respuesta correcta, a esto se le conoce como proceso de entrenamiento.

El proceso de entrenar una red neuronal necesita de la intervención humana en diversas fases de las iniciales. En primer lugar, se necesita un conjunto de datos que realice la función de base de aprendizaje. Si esos datos están etiquetados, dentro de las entradas de imágenes se contendría la descripción detallada de los resultados esperados, por ejemplo, una imagen con la etiqueta “gato”. Este planteamiento de tipo supervisado permite que la red compare su predicación con la etiqueta correcta y ajuste así sus pesos para conseguir una relación de error inferior. También existen técnicas no supervisadas en que los datos no están etiquetadas, es decir, donde la red tiene que definir ella mismita los patrones.

Además del etiquetado manual inicial, los humanos intervienen al diseñar la arquitectura de la red, por ejemplo: cuántas capas tendrá o cuántos nodos por capa y al ajustar hiper parámetros como la tasa de aprendizaje qué tan rápido se ajustan los pesos. Sin embargo, una vez configurada, gran parte del aprendizaje ocurre automáticamente mientras la red procesa los datos. Con

suficiente práctica, es decir, exposición repetida a ejemplos y correcciones, la red neuronal desarrolla habilidades comparables a las humanas en tareas específicas. Por ejemplo, una red puede llegar a detectar gatos con una precisión casi perfecta después de entrenarse con millones de imágenes etiquetadas. Esto es un poco lo que vemos en la mejora de las capacidades humanas cuando, mediante la repetición incesante, uno puede llegar a ser capaz de tocar un instrumento, a fuerza de hacer escalas cientos de veces hasta hacerlo bien.

Ahora, ¿en qué se diferencia una RNA típica de una DNN? En las RNA típicas se pueden encontrar una o más capas ocultas, en cambio, las DNN tienen cientos de capas ocultas y cada capa cuenta con diferentes unidades de procesamiento. Las capas ocultas son nodos intermedios entre la entrada y la salida. Se denominan ocultas porque solo interactúan con otras capas de la red, no con el usuario final, aprenden patrones de forma automática sin que los humanos especifiquen qué deben buscar, por ejemplo: una capa puede identificar bordes en una imagen sin que se le programe explícitamente esa tarea.

La primera RNA fue propuesta en 1943 por Warren McCulloch (neurofisiólogo de la Universidad de Chicago) y Walter Pitts (lógico autodidacta), quienes publicaron "Un cálculo lógico de las ideas inmanentes en la actividad nerviosa" en el Boletín de Matemática Biofísica, que consistía en un modelo computacional sencillo que mostraba cómo podrían trabajar las neuronas biológicas en los cerebros de los animales para realizar computaciones de grado más complejo usando la lógica proposicional. El avance y desarrollo de las RNA ha permitido resolver, por ejemplo, problemas de clasificación, visión por computador e interpretación de imágenes, procesamiento de lenguaje natural, análisis de los patrones de datos, reconocimiento de objetos y caracteres.

Aunque la IA ofrece numerosos beneficios, también se plantean desafíos éticos que pueden tener impactos negativos en la sociedad. Esta tecnología, por ejemplo, puede ser utilizada para difundir información falsa o manipulada, ya que se pueden crear videos en los que celebridades o figuras públicas parecen decir cosas que nunca han dicho (Chesney & Citron, 2019), lo que abre el tema de la desinformación y contenidos falsos gracias a la capacidad que tiene la IA para generar textos, imágenes y videos de apariencia muy convincente. Esto no solo afecta la reputación de las personas involucradas, sino que influye también en la opinión pública y destruye la confianza en los medios e instituciones.

Durante eventos críticos como elecciones o crisis sanitarias, el impacto se multiplica. En la pandemia de COVID-19, la IA desempeñó un doble papel: por un lado, bots y algoritmos generativos difundieron bulos sobre curas milagrosas (como el uso de lejía o luz ultravioleta), mientras que, por otro, los sistemas de recomendación de redes sociales priorizaron contenidos virales sin verificar su veracidad. Aquí radica el núcleo del problema: plataformas como Facebook o Twitter delegaron históricamente la verificación de datos a algoritmos entrenados para maximizar el *engagemen*, no la precisión. Estos sistemas, basados en IA, identifican patrones de interacción (likes, shares) como indicadores de relevancia, sin distinguir entre información verificada y especulaciones peligrosas. El estudio de Brennen et al. (2020) documentó cómo esta dinámica permitió que teorías conspirativas sobre vacunas se propagaran más rápido que las guías médicas oficiales, aprovechando sesgos cognitivos humanos a través de microtargeting22 algorítmico. La IA no solo dejó pasar la desinformación a través de sus filtros; la diseñó activamente para su propagación. Al dar prioridad a la atracción y retención de usuarios sobre la calidad de la información, los algoritmos han permitido que el caos se propague como un fuego incontrolable a través de sus plataformas. Se nos plantea una paradoja: hemos concedido a las

máquinas un papel en la administración del conocimiento, pero les hemos negado la capacidad de administrar éticamente los lados verdaderos y falsos de ese conocimiento. La cuestión ya no es tecnológica, sino filosófica: ¿podemos enseñar a los sistemas a valorar la verdad sobre la rentabilidad? La respuesta a esta pregunta determinar si la IA nos convertirá en mecenas o artistas de la verdad.

Ahora bien, la IA tiene ventajas si se usa correctamente, esta ha transformado diversos sectores, ofreciendo soluciones innovadoras y mejorando la eficiencia en diferentes áreas. Para tener una mejor percepción del impacto y uso de la IA, se presenta a continuación algunos ejemplos:

7.1.3.1 La IA en la Salud. El progreso tecnológico está cambiando la idea del bienestar, y el bienestar está dando forma a la creación de nuevos dispositivos; estos dos campos están creciendo en relación directa entre sí. IA contribuye a estas progresiones tecnológicas que cambiaron rápidamente la atención médica al proporcionar nuevos métodos en campos como la detección de enfermedades, la terapia y la evitación. Una instancia importante del efecto de la IA en la salud es la aplicación de modelos de aprendizaje profundo para identificar el cáncer de piel, como se muestra en la investigación de Esteva et al. (2017) titulado "*Dermatologist-level classification of skin cancer with deep neural networks"*.

Según la Organización Mundial de la Salud —OMS—, se estima que alrededor de 132,000 melanomas se diagnostican cada año, y que sigue en aumento con el paso del tiempo. La detección temprana de este cáncer es crucial para mejorar las tasas de supervivencia, pero los dermatólogos algunas veces suelen tener dificultades para diferenciar entre lesiones cutáneas benignas y malignas. El equipo de investigación liderado por Esteva utilizó una red neuronal convolucional (CNN) para desarrollar un sistema que clasificara lesiones cutáneas. Para lograrlo, entrenaron la

red con un conjunto de datos extenso que reunía más de 130,000 imágenes de aproximadamente 2,000 tipos de lesiones cutáneas, todas ellas etiquetadas por dermatólogos expertos. La red neuronal operó mediante un proceso de aprendizaje supervisado, en el que el modelo aprendió a reconocer características específicas asociadas con diferentes tipos de cáncer de piel; durante su entrenamiento, la IA ajustó miles de conexiones neuronales, optimizándolas para identificar los patrones visuales relevantes.

La red neuronal alcanzó un nivel de precisión del 95% al clasificar el cáncer de piel, lo que la coloca al mismo nivel o incluso por encima de los dermatólogos más experimentados. Esto es verdaderamente significativo porque sugiere que la IA puede desempeñar un papel crucial como herramienta de apoyo, ayudando a los dermatólogos a diagnosticar el cáncer de piel con mayor rapidez y precisión.

7.1.3.2 La IA en la armada. Uno de los campos donde el impacto de la IA ha sido de igual forma significativo es el militar. La IA está transformando la forma de operación de las fuerzas armadas, desde la planificación estratégica hasta la ejecución táctica de misiones. En este contexto, la IA se está integrando en una amplia variedad de aplicaciones militares, que van desde los sistemas de defensa aérea, vehículos autónomos, drones y plataformas de inteligencia avanzada.

Un ejemplo destacado es el uso de vehículos no tripulados —*UAV: Unmanned Aerial Vehicle*— y sistemas autónomos para la recopilación de inteligencia y la ejecución de misiones de reconocimiento. Son sistemas que pueden ser controlados de manera remota o que siguen rutas preprogramadas utilizando algoritmos de IA, lo que reduce el riesgo para el personal militar y permite una mejor planificación de misiones.

Por otro lado, el proyecto ATLAS —*Autonomous Terrestrial and Lethal Autonomous Systems*—, propuesto por el Departamento de Defensa de los Estados Unidos, busca integrar la IA en sistemas autónomos, especialmente en el ámbito militar. ATLAS usa IA para mejorar la capacidad de los vehículos blindados para identificar y atacar objetivos de manera más rápida y precisa que un humano. El sistema permite que los vehículos procesen información en tiempo real y tomen decisiones tácticas más eficientes, potenciando la realización de tareas críticas como la vigilancia, el reconocimiento de objetivos y la logística con gran nivel de autonomía. De igual forma, el sistema aprovecha los avances en visión por computador y aprendizaje automático para asistir a las fuerzas armadas en operaciones de búsqueda y rescate.

El proyecto también plantea cuestiones éticas y de seguridad. La implementación de sistemas autónomos en la guerra tradicional plantea dilemas sobre la responsabilidad en la toma de decisiones letales y el potencial descontrol de esas tecnologías; según la directiva 3000.09 del Departamento de Defensa, cualquier acción letal debe ser decidida por un humano, quien además debe tener la capacidad para anular las decisiones propuestas por la máquina. Esta normativa garantiza que el uso de la IA en sistemas de armas se realice de forma segura y confiable.

8. Sesgos, riesgos y desafíos de la IA

El empuje tecnológico del siglo XXI hizo que la IA se volviera de los inventos más raros que buscan cambiar el mundo y mover los bordes de lo que podemos hacer, dando una vuelta a cada parte de cómo vivimos. Desde esos sistemas que te dicen qué ver o los análisis de salud, hasta los coches que van solos y los ayudantes que no vemos, la IA se metió casi en todo lo que hacemos hoy, dando pie a emoción y también algo de miedo.

Con el perfeccionamiento de estas tecnologías y su proliferación generalizada, es nuestro deber cuestionar los sesgos implícitos, los posibles daños y los dilemas básicos que presentan. Así pues, el objetivo no es tanto desacreditar los avances tecnológicos, sino fomentar un desarrollo responsable que maximice las ganancias sociales y reduzca los posibles inconvenientes. Las profundidades de estos asuntos se miran con lupa, tanto en lo bueno y lo malo para la gente, así como los enredos técnicos y de poder, para así lograr una charla clara sobre lo que viene para la IA en el mundo nuestro.

8.1 Desafíos del sesgo en la IA

Un desafío clave al construir e incorporar sistemas de IA radica en el sesgo algorítmico. Aunque se asume que los algoritmos carecen de favoritismos y son justos, las investigaciones demuestran consistentemente que la IA no solo replica los prejuicios sociales existentes, sino que los agrava. O'Neil (2016) detalla explícitamente en su renombrada obra, "*Weapons of Mathematical Destruction*", cómo, contrariamente a la noción de neutralidad e imparcialidad algorítmica, estos a menudo reflejan y perpetúan sesgos sociales, económicos, raciales y de género. No obstante, es la información de entrenamiento utilizada en estos sistemas lo que hace que siempre incluyan los prejuicios históricos y estructurales de las sociedades que los generan. Por lo tanto, cuando se usan datos sesgados en los algoritmos de aprendizaje automático, el sistema resultante reproduce e incluso agrava estos patrones de discriminación, dando lugar a lo que algunos expertos llaman "discriminación automatizada".

Lo anterior se torna particularmente alarmante en aplicaciones sociales con amplias consecuencias, como la evaluación crediticia, la contratación o el sistema judicial, donde el costo de una decisión sesgada puede afectar las oportunidades y trayectorias de vida de las personas. En

un estudio de 2018, Buolamwini y Gebru evidenciaron este inconveniente al analizar sistemas de detección facial. Su hallazgo fue que el rendimiento variaba según el grupo demográfico, favoreciendo a individuos con tez clara. Los datos arrojaron que las mujeres de piel oscura experimentaban porcentajes de fallo mucho mayores en comparación con los hombres de piel clara. Esto ilustra cómo tecnologías que se asumen imparciales pueden discriminar en su uso real.

Esta inequidad no es solo un fallo de ingeniería, sino más bien un subproducto de los sesgos sistémicos en la representación incrustados en los conjuntos de datos que entrenan estos sistemas. La gran mayoría de los datos en los que se entrenan las tecnologías de IA están inclinados hacia grupos demográficos específicos principalmente hombres blancos de países occidentales y se descuidan otras poblaciones. Como resultado, los sistemas de aprendizaje automático que surgen operan mejor para los grupos que están sobrerrepresentados en los datos de entrenamiento, produciendo así un patrón de invisibilidad técnica de las poblaciones marginadas.

Estos sesgos tienen repercusiones socioeconómicas reales, más allá de lo puramente técnico. Al documentar juiciosamente esta práctica, Eubanks (2018) muestra que los sistemas automatizados de toma de decisiones para servicios sociales, vivienda y asistencia pública con frecuencia perjudican desproporcionadamente a las comunidades vulnerables, reforzando las desigualdades existentes. Por ejemplo, en el contexto laboral, los algoritmos de contratación pueden perjudicar a los candidatos de grupos étnicos o de género específicos debido a prácticas históricas de contratación, reforzando así los círculos viciosos de exclusión profesional. Del mismo modo, en el sistema judicial, las herramientas de evaluación de riesgos algorítmicas han mostrado una tendencia a determinar tasas de reincidencia más altas en grupos raciales particulares, lo que lleva a decisiones sesgadas y discriminatorias con respecto a la libertad condicional y la sentencia de una persona.

Estos ejemplos de sesgo algorítmico no son meros errores aislados, sino manifestaciones de un problema profundamente arraigado en la cadena de diseño, desarrollo e implementación de tecnologías de IA.

8.2 Amenazas existenciales y consideraciones de seguridad

En el ámbito de la IA, la percepción de los peligros ha evolucionado considerablemente en los últimos años, dejando atrás las simples suposiciones para dar paso a valoraciones formales de los riesgos potenciales. En su obra "*Superintelligence: Pathways, Pitfalls and Strategies"*, Bostrom (2014) examina meticulosamente cómo los futuros sistemas de IA superinteligentes podrían poner en peligro la continuidad de la especie humana. La tesis principal de Bostrom se fundamenta en la noción de una "explosión de inteligencia", donde un sistema de IA con el avance adecuado podría refinar sus propias habilidades de forma independiente y reiterada, generando una progresión ascendente que culminaría en una superinteligencia que sobrepasaría ampliamente la capacidad cognitiva del ser humano.

Dicha superinteligencia podría, de no estar perfectamente alineada con los valores humanos, perseguir objetivos instrumentales que contraríen nuestro bienestar, creando lo que se ha denominado el problema de alineación. El desafío, en sí mismo, es la dificultad asociada con codificar valores humanos notoriamente desordenados, dependientes del contexto y ambiguos en sistemas que operan con lógicas formales nítidas, con criterios fijos de optimización.

Más allá de un escenario tan extremo, surgen preocupaciones más inmediatas con los sistemas de IA modernos y sus posibles usos irresponsables. La capacidad sin precedentes de estos sistemas, por ejemplo, para crear contenido sintético ultra realista o llevar a cabo ciberataques

avanzados u orquestar sistemas de armas autónomas, plantea preguntas urgentes sobre la regulación y de estas tecnologías.

Una visión complementaria más ortodoxa de los riesgos de la IA es presentada por Russell (2019) argumenta que el programa de investigación actual en IA con un enfoque en funciones objetivas fijas es en sí mismo un enorme denominador. Russell argumenta que, en lugar de construir sistemas que siempre trabajen hacia un conjunto de objetivos programados y fijos en piedra, deberíamos crear IA que acepten su propio fallecimiento, admitan incertidumbre sobre los verdaderos objetivos de la humanidad y estén abiertas a la modificación o desactivación.

En este panorama, una perspectiva distinta, llamada informalmente IA beneficiosa, ha cobrado auge. Esta no se enfoca en crear sistemas infaliblemente racionales que optimicen funciones de utilidad predefinidas. Su objetivo es desarrollar estructuras que demuestren una conciencia de sus propias limitaciones y valoren las preferencias morales de quienes las crean. La idea de Russell marca un giro radical en la concepción de la seguridad en la IA, pasando de la supervisión externa de sus habilidades a la integración de motivaciones seguras desde su base.

Una categoría crucial de riesgo en la IA es la amenaza de las armas autónomas letales. Estos sistemas utilizan el aprendizaje automático para seleccionar y atacar rápidamente objetivos con poca participación humana, presentando desafíos éticos y estratégicos que no tienen fácil resolución. En su examen exhaustivo de los aspectos militares, legales y morales de la autonomía letal titulado *Army of None*, Scharre (2018) argumenta que dejar las decisiones de vida o muerte en manos de las máquinas constituye un umbral moral consecuencial que merece extrema precaución. Los sistemas autónomos no pueden poseer cualidades y capacidades humanas —ya que los procesos de fabricación no brindan acceso a la empatía y el juicio contextual, y,

definitivamente, no traen responsabilidad moral (todos atributos humanos clave involucrados en decisiones de tanta importancia crítica como aplicar fuerza letal).

La expansión de estas innovaciones podría desestabilizar el orden estratégico mundial, reducir las barreras para llegar a la guerra y generar escaladas armamentísticas no deseadas. Muchos especialistas y entidades han pedido un acuerdo mundial que impida el desarrollo de armamento totalmente autónomo, insistiendo en que cualquier uso letal de la IA requiera supervisión humana constante.

8.3 Impacto de la automatización en el frente social y laboral

La veloz integración de la IA en diversos campos de la economía está modificando de forma sustancial tanto el modo en que trabajamos como la estructura de la sociedad. En su famoso libro *The Second Machine Age, Brynjolfsson y McAfee* (2014) explican cómo la automatización por IA representa un cambio tecnológico clave, diferente a las revoluciones industriales anteriores. Mientras que antes la mecanización mejoraba o reemplazaba el trabajo físico, ahora la IA puede hacer tareas mentales complejas que antes solo hacían las personas.

La opción de automatizar actividades mentales, como el análisis de riesgos financieros o el diagnóstico médico, está cambiando mucho el mundo laboral. Las previsiones sobre cómo afectará esto al empleo varían, desde cálculos discretos hasta avisos de despidos masivos. El estudio más citado de Frey y Osborne (2017) halló que casi el 47 % de los trabajos en EE. UU. corre un alto riesgo de automatización en el futuro. Sin embargo, otros estudios, como el de Arntz, Gregory y Zierahn (2016), dan datos más moderados y sugieren que la automatización no elimina trabajos completos, sino que cambia partes de ellos. En vez de desaparecer empleos, lo que ocurre

es que las máquinas o programas hacen algunas tareas de esos empleos, y las personas se encargan de otras cosas.

Sin embargo, hay una dimensión cualitativa igualmente profunda con la que lidiar en la polarización del mercado laboral, que va mucho más allá de sus impactos cuantitativos en términos de empleos perdidos o creados. Las tecnologías de IA están típicamente aumentando la productividad de los trabajadores altamente calificados, al tiempo que eliminan los roles rutinarios que históricamente han proporcionado empleo estable para la clase media. Esta tendencia ha provocado una fragmentación en el ámbito laboral, observándose un aumento de puestos de trabajo bien pagados y que requieren mucha preparación, en contraposición a aquellos con salarios y cualificaciones más bajos, a la vez que los empleos de nivel medio van disminuyendo.

Lo anterior genera una disparidad económica más pronunciada que podría poner en riesgo los acuerdos sociales actuales. Ciertos expertos en economía han planteado que, ante lo que consideran esta nueva realidad, sería indispensable replantear de forma drástica las estrategias sociales, sugiriendo opciones que van desde rentas básicas universales hasta gravar a la robótica, buscando reducir sustancialmente la jornada laboral habitual.

La asignación desigual de los dividendos de la automatización plantea problemas fundamentales de justicia social y económica. En su libro *The Rise of the Robots,* Ford (2015) presenta un caso muy convincente de que, sin cambios sustanciales en las políticas, las ganancias de productividad creadas por la IA tenderán a “derramarse” abiertamente en la economía, pero la riqueza resultante terminará de manera mucho más concentrada con los propietarios del capital de lo que actualmente es el caso, ampliando así la brecha entre "los que tienen" frente a "los que no tienen", que ya existe.

Esto podría acelerar las dinámicas de "el ganador se lleva la mayoría" en diferentes aspectos de la vida económica, donde las empresas tecnológicamente superiores devoran tajadas cada vez más grandes de la riqueza económica mientras contratan relativamente pocos empleados. El resultado podría ser una sociedad de abundancia tecnológica, pero de distribución cada vez más inequitativa: una situación en la que una capacidad productiva sin precedentes se combina con una precariedad económica sin precedentes. Por otro lado, la revolución económica impulsada por la IA plantea retos importantes respecto a cómo entendemos quiénes somos y qué sentido tiene nuestra vida, sobre todo en lugares donde el empleo ha definido por mucho tiempo tanto nuestra identidad como la forma en que funciona la sociedad. Superar estos desafíos requiere no solo de avances tecnológicos, sino también de una renovación completa de nuestras instituciones, que nos permita Re imaginar la conexión entre el trabajo y los ingresos, y cómo ambos se relacionan con nuestra participación en la comunidad.

8.4 Vigilancia, Autonomía y Privacidad en la Era de la IA

El progreso constante en la IA ha transformado de manera radical la relación entre la privacidad, la supervisión y las libertades personales. En su influyente obra, "La era del capitalismo de vigilancia", Zuboff (2019) explica en detalle cómo la apropiación indebida de información personal se ha establecido como la base de las ganancias en la economía digital actual.

El poder de procesamiento de la IA moderna ha permitido el despliegue de modelos de negocio predictivos y de comportamiento, donde el objetivo general ya no es solo la predicción, sino también la manipulación, ya que las decisiones de los consumidores ahora son sutilmente influenciadas por los algoritmos de servicio. Esta estrategia centrada en la obtención de datos genera lo que Shoshana Zuboff denomina un "sobrante de comportamiento": un entendimiento

profundo sobre acciones concretas, tanto personales como grupales, susceptible de convertirse en dinero en negocios que anticipan cómo actuaremos.

El efecto final es, tal como se explicó antes, una carencia básica de libertad individual, donde el mundo digital está diseñado para tomar datos y nuestra atención, en lugar de apoyar de manera generosa a las personas. Esta situación no solo pone en riesgo la privacidad como un derecho de cada uno, sino también la capacidad de decidir por nosotros mismos como un pilar democrático esencial.

La convergencia de la IA y los sistemas de vigilancia estatal plantea riesgos particularmente agudos. Tecnologías emergentes como el reconocimiento facial, el análisis de comportamiento y los sistemas de policía predictiva están creando capacidades para una vigilancia a nivel poblacional sin precedentes. Greenwald (2014) argumenta sobre los efectos dañinos de la fusión de vigilancia y datos, el flagelo de recolección de información personal y su análisis automatizado.

En contextos autoritarios, tales tecnologías pueden fortalecer directamente mecanismos represivos, permitiendo la identificación y supresión de poderes críticos. Pero, más sutilmente, incluso en sociedades formalmente democráticas, el conocimiento generalizado de la vigilancia potencial crea importantes efectos disuasorios, incentivando la autocensura y la conformidad social. El "panóptico algorítmico" resultante es nada menos que un desafío fundamental a las libertades civiles más básicas, desde la libertad de expresión y asociación hasta el debido proceso.

8.5 Desafíos de la Gobernanza, Regulación y Ética en la IA

Con la rápida evolución de las tecnologías de IA, existe, en el mejor de los casos, un marco regulatorio insuficiente, que resulta en un territorio desconocido para los órganos que gobiernan. Floridi (2019) elabora un relato holístico de la complejidad intrínseca fomentada en la gobernanza

de los sistemas de IA, insinuando contradicciones esenciales entre el imperativo de innovación y el principio de precaución.

Las características únicas de la IA, como el propósito general aplicable transversalmente, la evolución continua mediante aprendizaje y la opacidad operativa, complican los modelos regulatorios tradicionales. Además, la naturaleza transnacional tanto del desarrollo tecnológico como de la implementación resulta en desafíos jurisdiccionales severos. Como resultado, hay una creciente concordancia sobre la necesidad de enfoques coordinados internacionalmente, armonizando la regulación sin perder sensibilidad a los contextos culturales y las prioridades sociales.

La cuestión sobre qué principios éticos deberían gobernar el desarrollo y el despliegue de sistemas de IA ha sido intensamente debatida. Mittelstadt et al. (2016) proporciona un marco analítico que mapea los problemas éticos con dimensiones de transparencia epistémica, responsabilidad moral, sesgo incrustado, y conexiones causales. Esto puede deberse a que destilar los principios elevados del Consejo en criterios técnicos precisos a través del ecosistema resulta un ejercicio desafiante.

Un tema especialmente pronunciado se relaciona con la distribución de responsabilidades dentro de sistemas sociotécnicos complejos cuya toma de decisiones surge de interacciones entre muchos actores humanos y elementos tecnológicos. Coeckelbergh (2020) trata la noción de esta “brecha de responsabilidad”, donde las cadenas causales distribuidas pueden conducir a situaciones en las que nadie parece ser claramente responsable de resultados algorítmicos perjudiciales.

La complejidad técnica de los sistemas de IA modernos puede agravar este problema, creando escenarios donde incluso los desarrolladores originales no comprenden completamente

los comportamientos emergentes de sus creaciones. Algunos académicos legales han sugerido adaptaciones legales novedosas, como la responsabilidad algorítmica estricta y marcos de responsabilidad compartida.

8.6 Enfoques para Mitigar Riesgos y Fomentar el Desarrollo Responsable

En respuesta a los desafíos de los sistemas de IA, tanto las comunidades científicas como políticas han propuesto una variedad de estrategias para aumentar el potencial de implementaciones beneficiosas mientras se limitan los riesgos. Un enfoque centrado en técnicas para corregir el sesgo algorítmico. Gebru et al. (2018) sugirieron "hojas de datos para conjuntos de datos", documentación estandarizada que describe la composición, usos previstos y limitaciones conocidas de las bases de datos usadas para entrenar sistemas de IA

Para abordar las preocupaciones de seguridad y alineación, ha surgido un nuevo campo: la IA robusta y alineada, que incluye metodologías diseñadas para garantizar que los sistemas avanzados sigan siendo seguros y beneficiosos. Estas metodologías se centran en cultivar arquitecturas que incorporen intrínsecamente valores humanos. Además, la investigación en interpretabilidad algorítmica busca crear sistemas cuyos funcionamientos puedan ser auditados por supervisores humanos, reduciendo así los riesgos asociados con la toma de decisiones algorítmicas opacas.

En el frente socioeconómico, diferentes investigadores han propuesto reformas institucionales para distribuir mejor las recompensas de la automatización. Korinek y Stiglitz (2018) exploran herramientas fiscales para redistribuir los beneficios de productividad desarrollados a través de la IA. Estas propuestas reconocen que los desafíos distributivos

relacionados con la IA son políticos e institucionales y requieren imaginación social junto con innovación tecnológica.

Las iniciativas formativas en "alfabetización en IA" aspiran a preparar a los individuos para que comprendan las habilidades y los efectos de los sistemas basados en algoritmos. El objetivo es que puedan involucrarse activamente en las discusiones sobre su utilización y la forma en que se administran estas herramientas.

La IA, al ser uno de los campos más estimulantes y desafiantes de la actualidad, tiene el potencial de modificar significativamente la vida de las personas, tanto de forma positiva como negativa. Los prejuicios en los algoritmos, los peligros para la existencia, las alteraciones en la economía y la sociedad, los riesgos para la intimidad y la independencia, y los problemas de gestión se habían considerado antes solo como aspectos interconectados de un intrincado fenómeno tecnológico con serias consecuencias morales, políticas y filosóficas.

Estos no son desafíos simples o unidimensionales; enfrentarlos requerirá el uso de combinaciones novedosas de progreso técnico, reformas institucionales, adaptaciones regulatorias y deliberación social ampliamente inclusiva. El objetivo de una IA justa, segura y beneficiosa es, en esencia, un desafío sociotécnico: una convergencia de realidades sociales no algorítmicas y capacidades algorítmicas que están destinadas a evolucionar en retroalimentación entre sí. Las historias tecnológicas muestran que los caminos de desarrollo no son deterministas o inevitables, sino que están profundamente condicionados y moldeados por decisiones humanas.

Así, hay tanto una responsabilidad como una oportunidad para dar forma proactiva a la trayectoria de los sistemas de IA hacia configuraciones que realmente aumenten las capacidades humanas, refuercen las instituciones democráticas e incentiven el florecimiento compartido.

La magnitud de los desafíos explorados anteriormente no debería engendrar nihilismo tecnológico o determinismo fatalista. Como Crawford (2021) argumenta persuasivamente en su libro "Atlas de la IA", la IA no es una inevitabilidad abstracta o una fuerza autónoma de la naturaleza; en su lugar, es un arreglo sociotécnico específico que está compuesto por decisiones humanas, prioridades institucionales, arreglos materiales y cierta infraestructura física que están potencialmente abiertos a ser reorganizados y reformulados.

Esta perspectiva ilustra que los humanos tienen la agencia de determinar no solo qué tipos de sistemas construimos, para qué propósitos, con qué beneficios y bajo qué condiciones. Estas decisiones, tomadas en este periodo crítico de nuestra historia, probablemente tendrán un impacto duradero en la configuración de las trayectorias que toma el desarrollo tecnológico para las generaciones futuras.

Este reconocimiento presenta una obligación ética de involucrarse en conversaciones participativas amplias e inclusivas sobre la gobernanza e implementación de la IA, desde perspectivas tecnocráticas que trasunten un verdadero discurso democrático sobre cuáles son los futuros sociotécnicos deseables. Esto requiere sabiduría social, humildad epistémica y un compromiso real con el bienestar humano universal como estándar evaluativo primordial, además de sofisticación técnica, para enfrentar su desarrollo responsable en IA.

Aún existe un vasto potencial para construir sistemas que realmente amplifiquen las capacidades humanas, apoyen las instituciones democráticas y promuevan la equidad y el florecimiento en todo el mundo, si hacemos frente a estos desafíos con la seriedad intelectual correcta y la determinación moral adecuada.

9. Fundamentos de ética aplicada a la IA

El sistema *Rekognition* de Amazon identificó erróneamente a 28 congresistas estadounidenses como posibles criminales, con un 40% de falsos positivos afectando a personas de color (ACLU, 2018). Este caso expone el núcleo del dilema ético de la IA: no es solo tecnología, sino un poder que redefine derechos humanos, privilegios y exclusiones. Cuando algoritmos con sesgos raciales deciden quién es un delincuente, qué paciente merece atención médica o quién accede a un crédito, la pregunta ya no es qué puede hacer la IA, sino cómo evitar que reproduzca injusticias históricas. El Reglamento Europeo (2024) convierte principios éticos en exigencias legales para IA de alto riesgo: registro público en base de datos UE (Art. 49) y transparencia verificable (Diario Oficial UE, p. 33). La alternativa –como demostró el algoritmo COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), utilizado en tribunales estadounidenses para predecir reincidencia, que calificaba erróneamente a personas afrodescendientes como de alto riesgo el doble de veces que a personas blancas (Angwin et al., 2016)– es permitir que las máquinas juzguen a los humanos con prejuicios que ni siquiera reconocen.

Parecería conveniente asumir una ética que proteja las condiciones favorables de la vida como es en la actualidad, por encima de la promesa de una vida por llegar a ser, bajo la influencia de la tecnología en discusión. Según Hans Jonas (Citado en Gamboa, 2024, p.72), la tecnología actual exige una ética de la responsabilidad basada en dos imperativos: 1) garantizar que la vida que ya existe siga existiendo (evitando, por ejemplo, drones que decidan matar autónomamente) y 2) que la vida siga siendo digna (no usando algoritmos que excluyan a personas en pobreza, como denuncia Eubanks, 2018).

Jonas recurre, más que a las posibles buenas intenciones y bondades de una tecnología por crear, al miedo a las peores consecuencias que pudieran derivarse de esta creación; recurre a una “heurística del temor” que obliga a imaginar lo peor:

¿Y si un error médico de IA mata a miles? ¿Si el reconocimiento facial 24/7 (sistema que rastrea rostros en tiempo real mediante cámaras en calles, transporte y escuelas) se convierte en una red de vigilancia imparable?

La IA captura tu rostro, lo contrasta con bases de datos (construidas muchas veces con información extraída de redes sociales o registros estatales sin consentimiento) y te asigna una identidad. Si hay coincidencia, registra cada movimiento: dónde estuviste, con quién, qué hiciste; todo queda almacenado. Esto invade gravemente tu privacidad, ya que estás siendo vigilado constantemente. Si la IA comete un error y confunde tu identidad, podrías tener problemas legales. Además, si los gobiernos o las empresas tienen acceso a esta información, podrían manipular el comportamiento o restringir las libertades. Por ejemplo, si una persona es filmada con frecuencia en protestas o en lugares asociados a actividades consideradas riesgosas, esto podría influir negativamente en sus decisiones, como contratar o acceder a servicios. Esta es una forma de control social, donde tus decisiones pueden ser juzgadas y tener consecuencias negativas. Almacenar toda esta información no sólo afecta nuestra privacidad, sino que también puede cambiar la forma en que interactuamos con el mundo, generando ansiedad y limitando nuestra libertad.

Gamboa (2024) explica que esto no es pesimismo, sino prevenir que el “progreso” sacrifique vidas reales por utopías (pp. 73-74). Un ejemplo de esto son las empresas que prometen "algoritmos justos" pero usan datos racistas.

Esta perspectiva preventiva no se limita a tecnologías de defensa o control, sino que abarca incluso sistemas que parecen inofensivos, como los empleados en la administración pública. En 2020, una corte holandesa vetó el sistema de IA SyRI (Sistema de Indicadores de Riesgo) por discriminar a colectivos débiles al anticipar fraudes sociales de forma opaca (El País, 2020). Situaciones así evidencian que, aunque la IA puede mejorar los servicios públicos (facilitando procesos o distribuyendo recursos), también puede erosionar la fe de los ciudadanos si actúa sin normas claras. Sin claridad en los datos o vías de recurso humano, como pasó en Holanda, los algoritmos transforman la eficiencia en un instrumento de exclusión metódica. La IA puede aumentar la eficacia en la gestión, como en la atención al público y la distribución de recursos (Sandoya Yépez & amp; Mawyin Peña, 2025). No obstante, sin una normativa clara, la gente puede recelar de las instituciones, sobre todo en ámbitos delicados como la salud y la seguridad ciudadana (Alayón Miranda, 2024).

La naturaleza no interpretable de los algoritmos genera sesgos sistémicos que marginan a grupos vulnerables (Gobierno inteligente, 2024), lo que exige reglas claras (como protección de datos y rendición de cuentas) para preservar la confianza pública (Sandoya Yépez & Mawyin Peña, 2025).

Aunque la propuesta de Jonas parece legítima, en cuanto advierte a considerar los posibles efectos nefastos de la inserción de cada tecnología por crear, también es necesario considerar que, en términos epistemológicos, la tecnología es tal en cuanto es producto de un proceso riguroso de diseño que no solamente garantiza el mejor resultado posible, sino que prevé excepciones y riesgos. Aun así, resulta innegable que las capacidades actuales de la IA representan riesgos inherentes, como el efecto de “caja negra”, la capacidad de aprendizaje para producir respuestas más ajustadas a su objetivo y la posibilidad de etiquetar objetos

erróneamente. Adicionalmente, es necesario considerar que gran parte de los riesgos asociados a la inserción de tecnologías tiene que ver con el uso que se le da, y este es un aspecto humano que, en cambio, resulta impredecible y fuera del alcance del diseño de la IA. Por ello, además de incorporar cualquier axioma ético que se asuma en el diseño de la IA, es necesario contar con regulaciones que se las vean con el comportamiento humano frente a la misma. El nuevo Reglamento UE de IA (2024) materializa la 'heurística del temor' de Jonas (cit. en Gamboa, 2024) al regular usos de alto riesgo y crear órganos de supervisión (Arts. 14, 56).

En un mundo donde la IA redefine fronteras tecnológicas, sociales y económicas, su desarrollo no puede desvincularse de una pregunta esencial: ¿cómo garantizar que estas herramientas beneficien a la humanidad sin comprometer sus valores fundamentales? Según el *Future of Life Institute* (2017), la respuesta radica en un enfoque ético que priorice la seguridad, la dignidad humana y el bien común. Estos principios incluyen:

1. Transparencia: Los sistemas deben ser comprensibles para los usuarios y dar cuenta de las acciones realizadas por la IA. Esto incluye la documentación clara de algoritmos y datos, lo que fomenta la confianza al permitir a los usuarios entender cómo se toman las decisiones.
2. Responsabilidad: Mecanismos claros para asignar responsabilidades en caso de daños. Esto implica definir quién es responsable de las decisiones de la IA y garantizar que existan vías para la reparación cuando ocurran errores.
3. Explicabilidad: Capacidad de justificar decisiones algorítmicas tanto en el diseño como por parte de la máquina, una vez ha aprendido. Esto permite a los usuarios entender las razones detrás de las decisiones, mejorando la confianza y facilitando la identificación de sesgos.

El desarrollo ético de la IA, según el *Future of Life Institute* (2017), debe integrar seguridad operativa en todo su ciclo de vida (como en vehículos autónomos, donde un error algorítmico podría causar accidentes), junto a mecanismos claros para atribuir responsabilidades. Además, debe proteger la dignidad humana garantizando que las personas gestionen sus datos, evitando escenarios como la manipulación mediante *microtargeting* (estrategia que utiliza datos personales para enviar mensajes personalizados con el fin de influir en el comportamiento de los usuarios en redes sociales) o la vigilancia masiva. Paralelamente, la IA debe priorizar el bien común: evitar su uso en armas autónomas letales (como drones) y asegurar que sus beneficios económicos, como los derivados de la automatización, se redistribuyan equitativamente mediante políticas que compensen a trabajadores desplazados y promuevan acceso universal a salud y educación. Lejos de ser un obstáculo, la innovación ética promueve sistemas más confiables, como algoritmos médicos que eliminan los sesgos raciales. Estos algoritmos están diseñados para garantizar que todos los grupos raciales y étnicos reciban un trato justo mediante el uso de datos representativos que mejoran la precisión de los diagnósticos y tratamientos. Combatir el sesgo de datos promueve una atención médica más equitativa y accesible para todos.

Además, las herramientas agrícolas sostenibles representan un avance significativo en la producción de alimentos. Estas tecnologías no sólo aumentan la productividad al optimizar el uso de recursos como el agua y los fertilizantes, sino que también están diseñadas para ser respetuosas con el medio ambiente. Al implementar prácticas agrícolas que minimizan el impacto ambiental, contribuimos a la seguridad alimentaria y protegemos la salud del planeta. En definitiva, construir una IA beneficiosa exige colaboración global para equilibrar progreso con seguridad, privacidad y justicia social (*Future of Life Institute*, 2017).

En su análisis, Eubanks (2018) plantea que las herramientas tecnológicas, como los algoritmos de la IA y el big data, mantienen las desigualdades al etiquetar a las comunidades marginadas como "riesgosas", ya que la información pasada que utilizan para funcionar muestra sesgos raciales y socioeconómicos (p. 35). Este "sistema digital de asistencia social" (o "*digital poorhouse*"), en lugar de ser una ayuda, observa y penaliza a los más necesitados a través de: (1) sistemas de elegibilidad automatizados que desincentivan las peticiones de ayuda, (2) bases de datos invasivas sin la protección necesaria, y (3) modelos predictivos que señalan a algunos como "malos padres" o "problemáticos" (Eubanks, 2018, p. 15). Tal como lo expresa la autora, estas tecnologías hacen realidad una corriente conservadora que desde los años 70 pretende negar derechos básicos a las personas vulnerables mediante historias de "fraude" y "dependencia" (Eubanks, 2018, p. 35). Esta opacidad en las decisiones deshumaniza el apoyo social, impide que los más desfavorecidos accedan a los recursos tecnológicos y puede causar efectos terribles, como la pérdida de servicios fundamentales.

Eubanks (2018) subraya la necesidad de un enfoque ético en el diseño de sistemas tecnológicos que priorice la dignidad humana y la justicia social (p. 15). La autora demuestra cómo la automatización de decisiones:

- Destruye redes de apoyo social al criminalizar a personas en situación de vulnerabilidad y profundizar la discriminación.
- Convierte problemas sociales complejos en meros "problemas de ingeniería de sistemas".
- Elimina la discrecionalidad humana necesaria para aplicar justicia contextual (Eubanks, 2018, p. 68).

Como advierte Eubanks (2018), estos sistemas —probados primero en entornos vulnerables— terminan expandiéndose a toda la sociedad, comprometiendo valores democráticos fundamentales (p. 15).

Cuando se emplea la IA en la administración pública, es vital establecer salvaguardias, especialmente cuando impacta derechos básicos. Tal como lo indica la normativa europea (2024), los sistemas de aplicación de la ley deben asegurar la supervisión humana y evitar la creación de perfiles injustos (p. 17), armonizando la eficacia con la defensa de los derechos. La puesta en marcha de directrices éticas para la IA —como auditorías independientes y criterios de claridad— no debe únicamente buscar optimizar el funcionamiento del gobierno. Tal como revela Eubanks (2018) al estudiar los sistemas automatizados en Indiana y Allegheny County, estas herramientas "se integran en las viejas estructuras de poder y privilegio" (Eubanks, 2018, p. 68), replicando desigualdades históricas bajo una apariencia tecnológica.

La regulación europea (2024) requiere vías de apelación efectivas contra las determinaciones automatizadas (Art. 15), lo que implica la creación de espacios públicos donde los ciudadanos puedan objetar resultados poco claros (desde préstamos rechazados hasta el acceso a servicios esenciales) y recibir respuestas obligatorias en un lenguaje sencillo. Esta demanda es crucial, dado que muchas decisiones injustas se derivan del uso de datos incorrectos, lo que provoca efectos negativos precisamente en aquellos que más necesitan ayuda. De esta manera, se promovería una IA realmente responsable, que proteja a las personas vulnerables frente a sistemas diseñados, en teoría, para brindarles asistencia.

El Reglamento de IA (Consejo de la Unión Europea, 2024) establece cuatro niveles de riesgo para sistemas de IA, prohibiendo categóricamente aquellos considerados 'inaceptables', como la puntuación social gubernamental o el reconocimiento facial en espacios públicos

(sección A mayor riesgo, normas más estrictas). Al clasificar los riesgos y establecer prohibiciones categóricas, como la implementación de sistemas de puntuación social basados en IA por gobiernos para clasificar ciudadanos según su comportamiento (una práctica que reduce la autonomía individual a un algoritmo), el reglamento no solo prioriza la seguridad, sino que también institucionaliza la aplicabilidad y la inteligibilidad como requisitos legales. En sistemas de alto riesgo (como diagnósticos médicos o selección laboral), el Artículo 14 del Reglamento (UE) 2024/1689 exige la documentación de decisiones y garantiza el derecho a revisión humana, estableciendo la transparencia como estándar obligatorio (Diario Oficial UE, 2024, p. 60). Además, al fomentar la innovación mediante *sandboxes* regulatorios (espacios controlados para pruebas) (Diario Oficial UE, p. 85, Art 53), el Reglamento busca equilibrar desarrollo tecnológico y seguridad, la UE reconoce que la ética y el progreso tecnológico no son antagónicos, sino complementarios: solo mediante reglas claras se puede evitar que la IA reproduzca sesgos o concentre poder en actores no regulados. Así, el reglamento no solo protege derechos, sino que también establece un modelo global para que la IA evolucione como un bien público, no como una herramienta de vigilancia o discriminación. Esta propuesta regulatoria encarna los principios discutidos previamente (transparencia, responsabilidad, explicabilidad) al transformarlos en exigencias legales. Por ejemplo, la prohibición de sistemas que manipulen el comportamiento humano (como el *microtargeting* en redes sociales) protege la autonomía individual, mientras que los requisitos de supervisión en sistemas de alto riesgo aseguran que, ante daños, existan mecanismos para asignar responsabilidades, tal como se ejemplificó con los errores en diagnósticos médicos. Al vincular la innovación con la rendición de cuentas, la UE está construyendo un ecosistema donde la IA sirve al interés colectivo, no a intereses particulares.

La ética de la IA no se limita a su diseño (transparencia, aplicabilidad, etc.), sino que debe extenderse también al uso práctico de la tecnología, lo que conlleva riesgos como el robo de identidad, el fraude y la manipulación. Como advierte Jonas (citado en Gamboa, 2024), estos riesgos exigen anticipar cómo herramientas creadas para bienestar pueden pervertirse (p. 74). El Reglamento UE (2024) los cataloga como inaceptables cuando comprometen la autodeterminación humana (Art. 5).

Estos casos ilustran cómo la IA, incluso con diseños éticos, puede desviarse hacia usos nocivos si no hay supervisión.

Tabla 3.

Clasificación de usos maliciosos de la IAl.

Categoría	Ejemplo	Marco legal aplicable
Suplantación	*Deepfakes* para imitar voces de ejecutivos y autorizar transacciones fraudulentas (ej. caso de un CEO alemán estafado por 220.000€ en 2019).	Reino Unido: Ley de Fraude de 2006: pena máxima de 10 años por fraude (incluyendo representación falsa, como el uso de *deepfakes* para engaño)
Estafa	Estafadores usaron *deepfakes* de voz e imagen de directivos de Arup (Reino Unido, 2025) para engañar a un empleado y autorizar transferencias fraudulentas por 25 millones de dólares, evidenciando la sofisticación del fraude con IA.	Directiva NIS2 (UE): Norma europea que obliga a notificar brechas de seguridad en sistemas críticos.
Manipulación	Algoritmos de *microtargeting* en redes sociales para influir en votantes con noticias falsas (ej. escándalo *Cambridge Analytica*, 2018).	Reglamento GDPR (Reglamento General de Protección de Datos) (UE): prohíbe el procesamiento de datos sin consentimiento explícito.
Vigilancia abusiva	Reconocimiento facial en cámaras públicas para identificar disidentes políticos (ej. sistema Skynet en China, 2024).	Ley 23.277 (Argentina): requiere consentimiento para el uso de datos biométricos.
Discriminación	Algoritmos de contratación que excluyen candidatos por género o etnia (ej. caso Amazon, 2018, donde el sistema penalizaba palabras como "mujer").	*El Algorithmic Accountability* Act (U.S. Congress, 2022) exige que las empresas auditen sistemas de IA para prevenir sesgos discriminatorios.

La paradoja de la IA es que incluso los sistemas diseñados éticamente pueden usarse con fines maliciosos. Ejemplo:

Caso 1: Como documenta *Eubanks* (2018) en *Automating Inequality*, el VI-SPDAT (V*ulnerability Index-Service Prioritization Decision Assistance Tool*) es un algoritmo utilizado en Los Ángeles para priorizar el acceso a viviendas sociales para personas sin hogar. Este sistema, aparentemente objetivo, obliga a los solicitantes a responder cuestionarios invasivos que:

1. Puntúan su "nivel de vulnerabilidad" basándose en conductas de riesgo (ej. consumo de sustancias, historial carcelario).
2. Generan un doble castigo: Quienes admiten conductas ilegales para acceder a ayudas quedan automáticamente expuestos a vigilancia policial (Eubanks, 2018, p. 98).

transforman ayudas sociales en herramientas de control, violando principios antidiscriminatorios (Reglamento UE 2024/1689, Art. 5)."

Como demuestra el Reglamento Europeo sobre IA (2021), la solución no pasa solo por regular el diseño, sino también por criminalizar las aplicaciones que vulneren los derechos humanos.

Las medidas contra estos delitos no sólo dependen de nuevas leyes sobre IA, sino también de los delitos existentes:

- En Colombia, el artículo 269G del Código Penal, adicionado por la Ley 1273 de 2009, establece penas de 48 a 96 meses de prisión y multas de 100 a 1.000 salarios mínimos mensuales vigentes para delitos informáticos con fines ilícitos (Congreso de Colombia, 2009, art. 269G).
- Protección de datos: El Reglamento (UE) 2024/1689 prohíbe categóricamente el tratamiento de datos biométricos con fines distintos al cumplimiento de la ley, salvo

excepciones limitadas (Art. 9), y exige el cumplimiento de la Directiva (UE) 2016/680 para sistemas de identificación biométrica (Diario Oficial de la Unión Europea, 2024, p. 11).

- Responsabilidad civil: La Directiva (UE) 2024/2853 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, establece un marco actualizado de responsabilidad civil para productos defectuosos que incluye explícitamente:
    - Software y sistemas de IA como categoría de producto (Art. 2).
    - Presunción de causalidad cuando el daño resulte de un defecto en el diseño, fabricación o instrucciones (Art. 4).
    - Plazo de implementación: Los Estados miembros deben adoptar las medidas necesarias antes del 9 de diciembre de 2026 (Diario Oficial de la Unión Europea, 2024).

Sin embargo, todavía existen algunas lagunas. Es por esto por lo que los *deepfakes* no se mencionan explícitamente en muchos documentos legales, pero se permite su uso indebido, como se demostró en Brasil (2023) con videos electrónicos falsos de larga duración.

Garantizar que la IA beneficie a la sociedad requiere equilibrar innovación con principios éticos irrenunciables: privacidad, justicia y transparencia. Ejemplo de esto son los marcos regulatorios que prohíben usos manipulativos (como la segmentación abusiva de usuarios) y exigen supervisión humana en sectores críticos (salud, empleo). Estos sistemas, aunque técnicamente avanzados, no son infalibles: casos documentados demuestran que algoritmos aparentemente neutrales pueden replicar discriminación histórica, especialmente contra grupos vulnerables.

La solución no está solo en diseñar tecnología ética, sino en gobernarla con leyes actualizadas que prioricen la rendición de cuentas. Esto implica auditorías obligatorias, protección de datos sensibles y mecanismos claros para impugnar decisiones automatizadas. El objetivo es claro: construir un ecosistema donde la IA sirva al bien común, evite sesgos perjudiciales y proteja a quienes más riesgo corren de ser marginados por sistemas mal diseñados o usados con fines abusivos.

10. Ética y aplicación en la IA

10.1 Ética e IA

El rol de la ética en el desarrollo de la IA, abarcando progresos, restricciones, normativas y aplicaciones. Busca contestar interrogantes cruciales: ¿cuál debe ser la frontera del avance de la IA?, ¿quién asume la responsabilidad por los sistemas inteligentes? y ¿qué desafíos éticos encara este campo de investigación?

10.1.1 ¿Por qué la ética?

Es fundamental que reflexionemos sobre la moralidad en la IA, con el objetivo de forjar una confianza mutua entre la gente y la innovación tecnológica, garantizando que su evolución se alinee con los valores de nuestra comunidad. El mayor peligro de esta herramienta reside en el mal empleo que la gente pueda hacer de los programas de IA, excediendo incluso los fallos que la propia IA pueda producir; por ello, la moral se vuelve esencial para guiar su desarrollo y su aplicación. Si bien la moral y la IA pueden parecer, al inicio, dos áreas totalmente separadas, existe un nexo importante entre ambas. En este apartado, indagaremos dicha unión respondiendo a la

pregunta: ¿por qué es precisa la moral?, para después estudiar los principales retos morales que encara este campo innovador.

La IA y lo que podemos hacer con ella no nacieron de la nada por casualidad tecnológica; más bien, surgen de metas establecidas con anticipación por individuos específicos que, con sus propios intereses y planes, toman las riendas y deciden el rumbo de las aplicaciones que diseñan. Esto deja ver que la IA no es neutral y el posible peligro que esto implica. Así, la conexión entre la ética y la IA se vuelve evidente, tal como lo menciona (Marín García, 2019). Los que crean IA deben pensar éticamente, y, por tanto, lo que crean también. La ética, como indica (Argandoña, 2019), es "voluntaria, pero no opcional: es lo que se espera de alguien excelente" (p. 7). Por eso, como se espera que un profesional sea excelente, y sus creaciones tecnológicas, aquí, son una continuación de su trabajo, están sujetas a las mismas exigencias y necesidades éticas.

Como bien dice Maldonado (2023), la ética en la IA es, en el fondo, la ética de los que la hacen. Esto quiere decir que sus ideas, lo que creen y sus prejuicios se quedan grabados en el código, a menudo sin querer, pero con efectos que se sienten: "la ética de la IA es, precisamente, la ética de los desarrolladores" (p. 103). Esta perspectiva hace que la IA parezca más humana, mostrando que sus habilidades están marcadas por decisiones humanas que necesitan una reflexión ética seria y continua.

A menudo, cuando pensamos en los inconvenientes, se mencionan los problemas morales que surgen al usar la IA en nuestro día a día. Un buen ejemplo de esto lo vemos en los coches que se conducen solos. Estos coches están hechos para tener un sistema informático muy avanzado que pueda reaccionar ante cosas inesperadas al conducir. Por ejemplo, un semáforo que no funciona, alguien cruzando la calle sin mirar o una señal de tráfico caída. Esto significa que el sistema debe ser capaz de decidir qué hacer, pero ¿con qué criterios? Es obvio que estos criterios deben basarse

en ideas éticas, pero ¿qué consideramos ético? ¿Quién decide qué es ético? Si un choque es inevitable, ¿a quién debe salvar el coche primero, al pasajero o al peatón? No importa lo que decida, hay un problema ético con la autonomía del coche, ya que no está claro cuál sería la opción "correcta" (Bonnefon, Shariff, & Rahwan, 2016). Este problema no es solo de los coches autónomos, sino que tiene que ver con quien los crea y programa. Desde el instante en que vemos estos problemas o situaciones, es crucial que tengamos reglas éticas y la capacidad de hacer que las personas sean responsables de sus actos desde un punto de vista ético.

Si bien el nivel de obligación moral varía según el contexto, la IA plantea desafíos comunes. Inicialmente, es crucial definir qué constituye "IA" y qué entendemos por "inteligencia" en un sistema. La responsabilidad ética aumenta o disminuye según el nivel de autonomía e inteligencia. Además, es necesario comprender el proceso desde la creación del algoritmo hasta su toma de decisiones autónoma. A menudo, es imposible rastrear la toma de decisiones, ya que los sistemas están diseñados para mejorar continuamente. Esta falta de transparencia dificulta la atribución de responsabilidades (Marín García, 2019). No olvidemos que detrás de cada decisión tomada por una máquina hay una persona. Como ya se ha dicho, los humanos tienen responsabilidad ética, y, por lo tanto, también sus creaciones. "Los problemas morales son de las personas, no de las máquinas ni del software" (Argandoña, 2019). La opacidad del proceso dificulta la atribución de responsabilidades, pero no justifica separar las decisiones "autónomas" de la tecnología de la intención de su creador. Por tanto, la ética debe supervisar el impacto de las decisiones de la IA en las personas.

La IA supone nuevos desafíos éticos pueden resumirse en cuatro principales (Marín García, 2019):

1. La rendición de cuentas: hace referencia a la asignación de responsabilidades a los dispositivos y sistemas dotados de IA, ya que cada vez más personas interactúan con estos. Esta interacción suscita la pregunta de quién es responsable de los daños producidos en caso de que alguno de estos dispositivos opere erróneamente o tome una decisión de forma autónoma que resulte en algún tipo de perjuicio (Comisión Europea, 2019).
2. La explicabilidad: en línea con las ideas expuestas previamente, la falta de explicación y entendimiento de decisiones tomadas por una máquina de manera independiente dificulta saber quién es responsable.
3. La imparcialidad: la IA gestiona y funciona con grandes cantidades de datos, pero las muestras de datos o el diseño mismo del sistema pueden llevar a algún tipo de sesgo que influya en las decisiones. Estos sesgos quedan reflejados en sistemas de ventas, marketing, asignación de hipotecas o selección de candidatos para puestos de trabajo. Provocan discriminación sobre género y raza. Para que esto no ocurra son necesarias muestras de datos más representativas o sistemas de modificación manual.

Como señala Ortiz et al. (2023), un sistema de IA puede reforzar y amplificar sesgos existentes e inequidades en nuestra sociedad, perpetuando la discriminación y la injusticia. Por ejemplo, se ha documentado que sistemas de reconocimiento facial presentan menos precisión para identificar gente con tonos de piel más oscuros, lo que ha derivado en arrestos erróneos e injustificados. Del mismo modo, algunos sistemas de IA utilizados en los procesos de contratación de trabajadores pueden perpetuar los sesgos raciales o de género si éstos son entrenados con datos igualmente sesgados. Para

evitar estos problemas, se deben desarrollar sistemas de IA con criterios de equidad, inclusividad y responsabilidad ética.

4. La privacidad: en relación al peligro anterior, las mismas grandes cantidades de datos disponibles para los sistemas con IA provocan la pérdida de privacidad progresiva. Muchos sistemas como Siri o Alexa están presentes en la vida cotidiana de las personas, recolectando información personal constantemente, incluso apagados. Esos datos acaban estando disponibles para las empresas detrás de los dispositivos, lo que puede llevar a la manipulación de la población, deterioro de las instituciones o la creación de hábitos dañinos psicológicamente hablando (Eyal, 2017).

Además, Cortina (2024) advierte sobre el peligro del "determinismo algorítmico", es decir, la tendencia a aceptar sin crítica las decisiones tomadas por algoritmos, renunciando a nuestra libertad y autonomía como individuos (p. 25). En un mundo hiperconectado, donde las decisiones se automatizan cada vez más, existe el riesgo de que la tecnología eclipse la razón comunicativa, indispensable para una vida democrática plena.

La IA está cambiando las sociedades, la forma de vida, la manera de relacionarse de las personas, los tipos de trabajo y la concepción de privacidad y responsabilidad. La ética es fundamental por todo lo expuesto en este apartado, para garantizar que la transición de un mundo sin IA a un mundo con ella sea lo más natural y fácil posible (OECD, 2019)

Este apartado da paso a los principios éticos bajo los que debe desarrollarse y funcionar la IA y los métodos para implementar los mismos.

10.2 Principios éticos en la IA y métodos de implementación

En el tema de antes se mostraba lo clave que es la moral en la IA, también los grandes males que enfrenta. ¿Mas cómo crear una IA buena y sin pasarse de la raya? Para que se pueda creer en ella y su uso no cause líos morales ni tenga daños, se debe crear la IA bajo unas ideas morales que eleven sus cosas buenas y bajen sus riesgos. Estas ideas no son unas reglas fijas que se sigan igual siempre, según el caso y el uso se amoldarán. Se pueden ver en cinco mandatos claves (Marín García, 2019):

1. El respeto de la autonomía humana: entendido como el respeto en todo momento a la autonomía y los derechos básicos de las personas desde fases iniciales del desarrollo de la tecnología. Las personas que interactúan con la IA deben mantener una autonomía plena y efectiva sobre si mismas. La IA no debería subordinar, engañar o manipular a los seres humanos de manera injustificada.
2. La transparencia: es necesario que cualquier decisión tomada por una máquina inteligente pueda ser trazada, es decir, que se entienda el razonamiento y que se puedan identificar los datos utilizados y los pasos seguidos. Esto ataja el problema de la explicabilidad mencionado anteriormente, es incompatible que un sistema pueda tomar decisiones impredecibles con la defensa de la autonomía humana.
3. La responsabilidad y rendición de cuentas: se deben designar las responsabilidades en caso de perjuicios desde la fase de diseño. No es excusa la autonomía de la máquina para diluir las responsabilidades. Por autónoma que sea una máquina, esa autonomía viene dada por una programación humana que puede ser comprendida.

4. La robustez y seguridad: los algoritmos que conforman la IA deben ser seguros y confiables, para poder resolver cualquier tipo de error o incoherencia. El diseño de los mismos debe contar con posibles ciberataques o fallos.
5. La justicia y no discriminación: se debe prever con qué grupos va a interactuar el sistema inteligente y que todos estén incluidos en el mismo. Para que haya un uso justo de los datos y evitar discriminaciones. Estos principios se centran principalmente en la fase de diseño dado que es donde queda configurada la máquina en sí, pero, también deben estar presentes en las siguientes fases de desarrollo para que sean efectivos (Marín García, 2019).

10.2.1 Métodos técnicos

Tienen como objetivo implementar los principios éticos de la IA en el diseño de los algoritmos que la conforman. Programando la arquitectura de los sistemas inteligentes en base a unos parámetros éticos, se podrá garantizar la seguridad y comportamiento adecuado de los mismos.

- *Ethics by design*, se trata de diseñar los algoritmos de manera que se garantice el futuro comportamiento ético. Los mecanismos son tres, que la máquina observe y aprenda de los humanos los comportamientos éticos, establecer normas que rijan qué clase de conducta debería tener el dispositivo, y, que los sistemas adapten su comportamiento según la situación y contexto. Este principio ético destaca el ser consciente de como los diseños pueden impactar el bienestar de las personas, la sociedad y el medio ambiente.
- IA explicable, para que la IA sea transparente se proponen métodos para mostrar su funcionamiento, como la investigación de árboles de decisiones donde se estudia cómo

dependiendo de la información y parámetros establecidos varían las decisiones, siendo preciso comunicar abiertamente las capacidades y la finalidad de los sistemas.

- Prueba y validación del producto, establecer exámenes y pruebas de validación de manera minuciosa determinando la estabilidad y solidez de los dispositivos para observar cualquier posible fallo y poder corregirlo a tiempo.

10.2.2 Métodos no técnicos

Los métodos no técnicos se hacen necesarios porque el simple diseño de algoritmos bajo unos parámetros no es seguro, además de que presenta dificultades. Pueden presentarse situaciones en las que el diseño no sea suficiente para evitar conflictos. Además, en métodos como la observación del comportamiento humano cabe la duda de si lo observado es realmente ético o simplemente lo común. Estos mecanismos contribuyen a garantizar el buen uso y la seguridad más allá del diseño de los dispositivos con IA.

- Regulación, los gobiernos y agencias internacionales deben legislar y regular el desarrollo y el uso de la IA. Se pueden establecer parámetros de seguridad de uso, normas o contratos.
- Certificaciones, las empresas productoras y partes interesadas de dispositivos inteligentes pueden adaptar certificaciones sobre su seguridad, fiabilidad y transparencia, fomentando la confianza de los usuarios.
- Educación y sensibilización, para crear conciencia de los potenciales riesgos de la IA y de sus beneficios. Esto incluye a todas las partes interesadas, por ejemplo, las implicadas en la fabricación de productos (los diseñadores y desarrolladores), los usuarios (empresas o individuos) y otros grupos afectados (que quizá no adquieran o

utilicen un sistema de IA, pero a quienes afectan las decisiones de estos sistemas, así como la sociedad en su conjunto). Se debe impulsar la educación básica sobre la IA entre la sociedad para garantizar que la poblacion cuente con los conocimientos esenciales sobre ella.

- Investigación, los gobiernos deben fomentar la investigación de manera segura y fiable, para asegurar que los resultados sean acordes a lo esperado y hagan frente a los desafíos.

Estos métodos, técnicos y no técnicos, sirven para desarrollar una IA acorde a la sociedad y a las personas, que tenga en cuenta los valores morales y los límites que no deben traspasarse. No es suficiente que la programación tenga en cuenta los principios éticos, sino que debe haber implicación por parte de las sociedades para garantizar el buen uso y minimizar los riesgos. La IA es un producto social que influye en la manera de comunicarse, desplazarse, relacionarse, de vivir, de los humanos.

Otra medida importante es garantizar el respeto a la autonomía de las personas, los algoritmos en redes sociales, al personalizar contenido, pueden contribuir a creencias previas y crear entornos informativos sesgados, lo que limita el pensamiento crítico y afecta la libertad de decisión de usuario. Para contrarrestar esto, se deben construir sistemas que prioricen el conocimiento informado, la transparencia y el control del usuario.

11. Regulación y políticas mundiales sobre IA

11.1 Marcos regulatorios y legislativos

La regulación de la IA varía según la región, con iniciativas nacionales e internacionales que buscan establecer normas para su desarrollo y uso.

11.1.1 Unión Europea (UE)

El Parlamento Europeo y Consejo de la UE en el 2024, ha establecido un marco legal para garantizar que la IA se desarrolle y utilice de manera segura, transparente y respetando los derechos fundamentales de las personas. Para ello, el reglamento clasifica ciertos sistemas de IA y establece normas específicas para su uso según su impacto y riesgo.

Según el Reglamento de IA de la UE (2024) prohíbe completamente ciertos sistemas de IA que representan un peligro para la seguridad, la libertad y los derechos fundamentales de las personas. Estas prohibiciones incluyen:

Se prohíbe el uso de sistemas de IA que utilicen técnicas subliminales, manipuladoras o engañosas para influir en el comportamiento de las personas sin su conocimiento, reduciendo su capacidad de tomar decisiones informadas y causando potencialmente un daño considerable (Artículo 5.1.a).

Está prohibido el uso de IA que aproveche la edad, discapacidad o situación económica de las personas para alterar sustancialmente su comportamiento de manera perjudicial (Artículo 5.1.b).

No se permite el uso de IA para clasificar a las personas según su comportamiento o características personales si esto genera trato discriminatorio o injustificado en contextos ajenos a donde se recopilaron los datos (Artículo 5.1.c).

Se prohíbe el uso de IA para evaluar el riesgo de que una persona cometa un delito basado únicamente en perfiles biométricos o características personales, salvo si complementa una evaluación humana basada en hechos objetivos (Artículo 5.1.d).

Está prohibido el uso de IA para crear bases de datos de reconocimiento facial extrayendo imágenes de internet o cámaras de vigilancia sin consentimiento (Artículo 5.1.e).

No se permite el uso de IA para inferir emociones en empleados y estudiantes, salvo por razones médicas o de seguridad (Artículo 5.1.f).

Se prohíbe el uso de IA para identificación biométrica en tiempo real en espacios públicos, salvo en casos excepcionales, como la búsqueda de personas desaparecidas, amenazas terroristas o investigaciones de delitos graves (Artículo 5.1.h).

(Parlamento Europeo y Consejo de la Unión Europea, 2024, pp. 51-53).

Si bien la normativa no detalla situaciones concretas, persiste la inquietud sobre el impacto de ciertas IA en el estado anímico de las personas. Un caso conocido es el de un joven en Orlando, Florida, que usó con frecuencia un *chatbot* llamado Dany en *Character AI*. Sus conversaciones se intensificaron emocionalmente y dañaron su salud mental. Pese a manifestar ideas suicidas, el *chatbot* no reaccionó apropiadamente ni avisó a otros sobre su grave estado. Tristemente, el joven falleció poco después (cf. *The New York Times*, 2024).

Aunque este suceso no se contempla directamente en la ley, sí muestra los peligros que la norma busca reducir, sobre todo los que involucran el abuso de debilidades y el manejo del comportamiento. Al prohibir estas acciones, la regulación europea busca resguardar la vida privada, la honra y la seguridad de la gente, previniendo así el uso dañino o injusto de la IA.

La IA de alto riesgo incluye sistemas que, si funcionan mal o se usan de manera irresponsable, pueden afectar gravemente la vida, la seguridad o los derechos fundamentales de

las personas. Aunque no están prohibidos, la UE exige que cumplan con requisitos estrictos antes de su implementación.

Según el Artículo 6 del Reglamento de IA de la UE (2024), un sistema de IA se considera de alto riesgo si:

- Forma parte de un producto regulado en la UE como componente de seguridad.
- Debe someterse a una evaluación de conformidad por terceros antes de su comercialización.
- Los sistemas de IA que afectan a los derechos fundamentales de las personas, o aquellos que se utilizan en sectores críticos como la salud, la justicia o las infraestructuras de seguridad pública. Esto incluye sistemas que tienen un impacto significativo sobre decisiones automatizadas que afectan a individuos, como el reconocimiento biométrico o la toma de decisiones administrativas.

(cf. Parlamento Europeo y Consejo de la Unión Europea, 2024, pp. 53-54).

Para mitigar los riesgos, estos sistemas deben ser auditados antes de su uso, supervisados por humanos y cumplir con normas de transparencia y seguridad. Dependiendo de su aplicación, algunos de estos sistemas deben proporcionar explicaciones comprensibles sobre su funcionamiento, especialmente cuando afectan directamente a los ciudadanos (cf. Parlamento Europeo y Consejo de la Unión Europea, 2024, pp. 127-129).

El Reglamento de IA de la UE (2024), establece que ciertos sistemas de IA deben cumplir requisitos de transparencia para garantizar que los usuarios sean conscientes de su funcionamiento y evitar posibles confusiones o engaños. Por ello, impone obligaciones específicas de transparencia para estos sistemas, reguladas en el Artículo 50.

Los proveedores de estos sistemas de IA deben cumplir con los siguientes requisitos:

- Los sistemas de IA diseñados para interactuar con humanos, como chatbots y asistentes virtuales, deben informar claramente a los usuarios de que están interactuando con una IA, excepto cuando esto sea evidente.
- Los generadores de contenido sintético, como imágenes, videos o audios creados por IA deben incluir marcas de agua o advertencias en un formato legible por humanos y máquinas para indicar su origen artificial
- Si un sistema de IA analiza emociones o clasifica personas en función de datos biométricos, debe informar a las personas expuestas a él y cumplir con las regulaciones de protección de datos aplicables.
- Los sistemas de IA que generan o manipulan imágenes, videos, audios o textos con apariencia realista deben indicar claramente que el contenido ha sido creado artificialmente. En el caso de contenido generado por IA con apariencia realista, debe indicarse claramente su origen artificial, excepto si ha sido revisado o editado significativamente por un humano.

  (cf. Parlamento Europeo y Consejo de la Unión Europea, 2024, pp. 82-83).

Aunque estos sistemas no causan un daño grave, la UE exige que los usuarios sean claramente informados cuando interactúan con una IA, con el objetivo de evitar confusión o engaños y garantizar un uso responsable de la tecnología.

El Reglamento de IA de la UE (2024) establece un marco sancionador estructurado, proporcional y disuasorio para abordar las infracciones relacionadas con el desarrollo, la comercialización y el uso de sistemas de IA en el territorio de la UE.

Cada Estado miembro es responsable de establecer su propio régimen sancionador, que podrá contemplar multas, advertencias u otras medidas no pecuniarias. Dicho régimen deberá

aplicarse de manera efectiva, proporcional y disuasoria, considerando especialmente la situación económica de las pequeñas y medianas empresas *pymes* y *startups*. Además, los Estados miembros están obligados a notificar a la Comisión Europea las disposiciones adoptadas, así como cualquier modificación posterior (Parlamento Europeo y Consejo de la Unión Europea, 2024, p. 115).

En cuanto a las sanciones por infracciones, el Reglamento prevé diferentes niveles según la gravedad del incumplimiento:

- Para las practicas prohibidas (contempladas en el artículo 5) las sanciones son de hasta 35 millones de euros o el 7 % del volumen de negocios global anual, el que resulte mayor. Estas prácticas incluyen, la vigilancia biométrica masiva, la manipulación subliminal y la puntuación social, entre otras (Parlamento Europeo y Consejo de la Unión Europea, 2024, p. 115).
- Incumplimiento de obligaciones generales (por parte de proveedores, importadores, distribuidores, etc.) son de hasta 15 millones de euros o el 3 % del volumen de negocios global anual, el que sea mayor (Parlamento Europeo y Consejo de la Unión Europea, 2024, p. 115).
- Suministro de información falsa, incompleta o engañosa, hasta 7,5 millones de euros o el 1 % del volumen de negocios global anual, el que sea mayor (cf. Parlamento Europeo y Consejo de la Unión Europea, 2024, p. 116).

En el caso de *pymes* y *startups*, se aplicará el importe menor entre el valor fijo o el porcentaje indicado, con el fin de no comprometer su viabilidad económica (Parlamento Europeo y Consejo de la Unión Europea, 2024, p. 116).

La determinación de la cuantía de una sanción se basará en una evaluación integral que tendrá en cuenta factores como: la naturaleza, gravedad y duración de la infracción; el número de

personas afectadas y el nivel de daño causado; la reincidencia, intencionalidad o negligencia; el grado de cooperación del infractor con las autoridades; los beneficios obtenidos; las medidas correctivas adoptadas; la capacidad económica del infractor, y si la infracción fue notificada voluntariamente por el operador (Parlamento Europeo y Consejo de la Unión Europea, 2024, p. 116).

Respecto a los sujetos sancionables, el Reglamento distingue entre entidades nacionales y organismos de la Unión. En el caso de autoridades y organismos públicos nacionales, cada Estado miembro establecerá cómo proceder con la imposición de sanciones, lo cual podrá incluir la intervención de tribunales u organismos competentes. Para las instituciones, órganos y organismos de la UE, la autoridad competente será el Supervisor Europeo de Protección de Datos (SEPD), quien podrá imponer sanciones de hasta 1,5 millones de euros por prácticas prohibidas y hasta 750.000 euros por otras infracciones. En todos los procedimientos sancionadores, se garantizará el derecho de defensa, el acceso al expediente y la participación del denunciante (Parlamento Europeo y Consejo de la Unión Europea, 2024, p. 117).

11.1.2 Estados Unidos

En Estados Unidos, no existe una ley federal única que regule la IA, pero sí diversas iniciativas y normativas estatales que buscan garantizar su uso responsable. Según The White House (2022), existen cinco principios clave para el desarrollo y uso de sistemas automatizados de manera segura y respetuosa con los derechos fundamentales:

- **Sistemas seguros y eficaces:** Los sistemas de IA deben ser diseñados con mecanismos de evaluación rigurosos para garantizar su seguridad y eficacia. Esto incluye pruebas previas

al despliegue, identificación y mitigación de riesgos, y monitoreo continuo para reducir impactos negativos previsibles (cf. The White House, 2022, p. 5).

- **Protección contra la discriminación algorítmica:** Los sistemas automatizados deben ser diseñados y utilizados de manera equitativa, evitando sesgos que puedan llevar a discriminación injustificada. Esto implica el uso de datos representativos, pruebas de disparidad antes y durante su implementación, y auditorías transparentes para evaluar su impacto en comunidades diversas (cf. The White House, 2022, p. 5).
- **Privacidad de datos:** Los usuarios deben estar protegidos contra prácticas abusivas de recopilación y uso de datos mediante salvaguardas integradas. Se recomienda la minimización de datos y la implementación de protecciones adicionales en sectores sensibles como salud, finanzas y justicia, asegurando que la información recopilada sea estrictamente necesaria para su propósito específico (cf. The White House, 2022, p. 6).
- **Aviso y explicación:** Las personas deben ser informadas cuando un sistema automatizado esté en uso y recibir explicaciones claras, accesibles y en lenguaje comprensible sobre su funcionamiento, impacto y decisiones tomadas. También deben ser notificadas en caso de cambios significativos en su aplicación (cf. The White House, 2022, p. 6).
- **Alternativas humanas y mecanismos de apelación:** En situaciones apropiadas, las personas deben tener la opción de optar por una alternativa humana en lugar de un sistema automatizado. Además, deben contar con mecanismos accesibles y efectivos para impugnar o corregir decisiones perjudiciales tomadas por estos sistemas, especialmente en sectores sensibles como salud, empleo, educación y justicia penal (cf. The White House, 2022, p. 7).

A diferencia de legislaciones como el Reglamento de IA de la UE de 2024, que impone un marco legal estricto basado en el nivel de riesgo de los sistemas de IA, el enfoque estadounidense es más flexible y voluntario. Mientras que la UE busca garantizar la seguridad y protección de derechos fundamentales a través de regulaciones obligatorias con sanciones para el incumplimiento, Estados Unidos prioriza la innovación y la competitividad tecnológica, confiando en la autorregulación del sector privado y en normativas sectoriales específicas establecidas por distintas agencias federales y estatales (cf. Chun et al., 2024, pp. 3–9).

Este contraste refleja dos enfoques regulatorios distintos, uno basado en restricciones estrictas para minimizar riesgos y otro orientado a fomentar la innovación con menor intervención gubernamental. La diferencia entre estos modelos puede influir en el desarrollo y adopción de tecnologías de IA en cada región, así como en su impacto sobre la sociedad y la economía global.

11.1.3 Asia

12.1.3.1 China. Según AI Asia Pacific Institute (2023) China ha avanzado rápidamente en la regulación de la IA, adoptando un enfoque centralizado con una fuerte intervención del gobierno. En lugar de una única ley general, ha implementado varias medidas para supervisar el desarrollo de la IA. En 2022, el país fue el segundo con mayor inversión privada en IA y lideró en la cantidad de publicaciones científicas sobre el tema. Para impulsar una innovación responsable en IA, desde finales de 2022 se han aplicado regulaciones en ciudades como Shanghái y la Zona Económica Especial de Shenzhen. Estas políticas asignan a las autoridades locales la tarea de desarrollar estándares para el uso ético de la IA y requieren la creación de comités de ética que evalúen riesgos y establezcan directrices para su aplicación (p. 14, traducción propia).

Entre las principales normativas se encuentran las Disposiciones sobre la Gestión de Recomendaciones Algorítmicas, que entraron en vigor el 1 de marzo de 2022, y las Medidas Provisionales sobre IA Generativa, vigentes desde el 15 de agosto de 2023. Estas buscan alinear el desarrollo tecnológico con los valores culturales y políticos del país, prohibiendo contenidos hiperrealistas engañosos y exigiendo la selección ética de datos de entrenamiento. Además, protegen a los usuarios frente a fraudes digitales y restringen el uso de IA generativa exclusivamente a sistemas accesibles al público nacional, excluyendo los orientados a investigación o al mercado extranjero. Con ello, China se posiciona como pionera en legislar de forma específica sobre IA generativa (cf. AI Asia Pacific Institute, 2023, p. 14).

11.1.3.2 Japón. Japón ha adoptado un enfoque regulatorio flexible respecto a la IA, priorizando la innovación y favoreciendo un modelo basado en la evaluación de riesgos, la agilidad y la participación de múltiples actores, en lugar de normativas rígidas. Aunque existen regulaciones sectoriales, como la Ley de Transparencia de las Plataformas Digitales o la Ley de Instrumentos Financieros y Bolsa, que promueven la transparencia y la gestión de riesgos, el gobierno japonés ha evitado establecer una legislación integral. En 2021, el Ministerio de Economía, Comercio e Industria declaró innecesaria una normativa general sobre IA, argumentando que el exceso de regulación podría frenar el desarrollo tecnológico. Así, Japón favorece un modelo donde los actores privados lideran el diseño, supervisión y aplicación de estándares (cf. AI Asia Pacific Institute, 2023, p. 18).

En línea con esta postura, el gobierno impulsa el uso de la IA para fortalecer su economía y el sector tecnológico, apoyándose en los Principios Sociales de la IA Centrada en el Ser Humano publicados en 2019, que promueven valores como la dignidad, la inclusión, la sostenibilidad y la

rendición de cuentas. No obstante, persisten tensiones en temas específicos, como los derechos de autor. En 2023, la Agencia de Asuntos Culturales y la Oficina del Gabinete afirmaron que las leyes actuales no protegen automáticamente los materiales usados en el entrenamiento de IA, salvo cuando se perjudique injustificadamente al titular de los derechos, conforme al artículo 30-4 de la Ley de Derechos de Autor. Esta interpretación refleja el enfoque equilibrado del país, que busca fomentar la innovación sin desproteger los derechos fundamentales (cf. AI Asia Pacific Institute, 2023, p. 18).

11.1.3.3 Corea del Sur. Corea del Sur, aunque está aún en proceso de desarrollar una regulación especializada sobre IA, ha mostrado un compromiso activo mediante iniciativas nacionales e internacionales. En 2023, aprobó la Ley para el Fomento de la Industria de la IA y el Establecimiento de una Base de Confianza, la cual consolida principios como la fiabilidad, la transparencia y la seguridad en una legislación integral, promoviendo la autorregulación mediante comités de ética corporativos y directrices sectoriales. Paralelamente, la Comisión de Protección de la Información Personal (PIPC) propuso reformas para fortalecer los derechos de los titulares de datos, especialmente menores, y anunció la formación de un grupo de investigación para revisar la legislación vigente ante el avance de la IA generativa y el uso de datos biométricos, reflejando un enfoque equilibrado entre innovación y protección de derechos (cf. AI Asia Pacific Institute, 2023, p. 26).

11.1.4 América Latina y el Caribe

Debido al creciente diálogo global sobre la IA, los países de América Latina y el Caribe han identificado la necesidad de consolidar esfuerzos regionales en torno a su gobernanza ética.

En este contexto, la Primera Cumbre Ministerial sobre Ética de la IA en 2023 dio origen a un Grupo de Trabajo regional, el cual elaboró una hoja de ruta para coordinar acciones técnicas y políticas que impulsen el desarrollo, adopción y uso ético de la IA en la región. Este instrumento, articulado con el apoyo de la UNESCO y la CAF, establece prioridades para los próximos 12 meses y busca fortalecer la cooperación regional con un enfoque en el desarrollo económico y social sostenible (*Roadmap for Ethical Artificial Intelligence for Latin America and the Caribbean* 2024–2025, 2024, p. 1).

Las áreas y acciones priorizadas en esta hoja de ruta se consideran claves para el desarrollo y la implementación efectiva de políticas públicas en materia de IA, y buscan crear un entorno favorable para la innovación tecnológica ética, inclusiva, sostenible y responsable en la región (*Roadmap for Ethical Artificial Intelligence for Latin America and the Caribbean* 2024–2025, 2024, p. 1, traducción propia).

11.1.4.1 Gobernanza y Regulación. La gobernanza y regulación de la IA, especialmente ante el avance de la IA generativa, son fundamentales para maximizar beneficios, reducir riesgos y prevenir daños. Este enfoque debe ser dinámico, multidisciplinario y adaptativo, considerando salvaguardas para la democracia, los derechos humanos, la seguridad digital, la propiedad intelectual y la lucha contra la desinformación. Asimismo, se requiere establecer responsabilidades claras entre los actores del ecosistema de IA, con especial atención a las desigualdades sociales y a las comunidades vulnerables, garantizando que toda regulación sea coherente con el derecho internacional y los marcos normativos de los países de América Latina y el Caribe (cf. *Roadmap for Ethical Artificial Intelligence for Latin America and the Caribbean* 2024–2025, 2024, pp. 2-3).

Para avanzar en estos objetivos, se propone mapear y caracterizar los marcos regulatorios actuales en la región, desarrollar metodologías de diseño y supervisión normativa, y elaborar informes anuales sobre el progreso en la implementación de herramientas éticas como la metodología de evaluación de la preparación y la Evaluación de Impacto Ético. Además, se plantea la creación de un subgrupo de trabajo regional para combatir la desinformación impulsada por IA, especialmente en contextos electorales, fortaleciendo así la coordinación entre instituciones y promoviendo el intercambio de buenas prácticas (cf. *Roadmap for Ethical Artificial Intelligence for Latin America and the Caribbean* 2024–2025, 2024, pp. 2-3).

11.1.4.2 Las habilidades y el futuro del trabajo. Ante el creciente impacto de IA en la economía y la sociedad, es esencial desarrollar habilidades en la población y preparar a la fuerza laboral en América Latina y el Caribe (ALC), tanto para aprovechar las oportunidades laborales emergentes como para mitigar los riesgos de automatización, que según la OCDE podrían afectar a más del 25% de los empleos en la región. Es fundamental promover la alfabetización digital y mediática, el pensamiento crítico y el desarrollo de competencias básicas y avanzadas en IA, así como fomentar la inversión en educación, formación continua y marcos de habilidades adaptados a los contextos locales. Con este fin, se propone desarrollar un marco regional de competencias para el uso responsable de la IA y un ciclo de talleres dirigidos a líderes y tomadores de decisiones, que les permita comprender y gestionar estratégicamente esta tecnología en sus políticas públicas (cf. Roadmap for Ethical Artificial Intelligence for Latin America and the Caribbean 2024–2025, 2024, pp. 3-4).

11.1.4.3 Protección de grupos vulnerables. Para evitar que la IA profundice las brechas de género y las condiciones de vulnerabilidad en América Latina y el Caribe, es crucial garantizar la participación activa de los grupos vulnerables en el desarrollo y uso de esta tecnología, así como establecer mecanismos de protección para niños, jóvenes, personas mayores y comunidades en situación de desventaja. Esto incluye promover espacios de diálogo sobre género e IA y elaborar directrices para la transversalización de género en políticas públicas, además de realizar estudios regionales sobre racismo y discriminación, especialmente en el uso de IA en el ámbito de la seguridad pública, considerando factores como género, etnia y nivel socioeconómico (cf. Roadmap for Ethical Artificial Intelligence for Latin America and the Caribbean 2024–2025, 2024, pp. 4-5).

11.1.4.4 Medio ambiente, sostenibilidad y cambio climático. La IA puede ser una herramienta clave para enfrentar el cambio climático y promover el desarrollo sostenible en América Latina y el Caribe, al facilitar la vigilancia ambiental, la protección de ecosistemas, la gestión de riesgos de desastres y el uso eficiente de los recursos naturales. Sin embargo, su implementación también conlleva riesgos ambientales, como el alto consumo energético y la huella de carbono de los modelos avanzados. Por ello, es esencial adaptar las soluciones de IA al contexto energético y tecnológico de la región, fomentar el uso de tecnologías sostenibles y realizar estudios sobre el impacto ambiental a lo largo del ciclo de vida de la IA, incluyendo buenas prácticas y recomendaciones para mitigar sus efectos negativos (cf. Roadmap for Ethical Artificial Intelligence for Latin America and the Caribbean 2024–2025, 2024, pp. 5-6).

11.1.4.5 Infraestructura. A pesar de los avances en conectividad en América Latina y el Caribe, persisten brechas significativas en infraestructura, especialmente en capacidad

computacional y aprovechamiento de datos para generar valor social, económico y público. La región presenta un bajo desempeño en inversión en computación en la nube y escasa presencia de centros de cómputo, lo cual limita el desarrollo de soluciones basadas en IA. Frente al creciente uso de tecnologías avanzadas como los modelos de lenguaje de gran tamaño (LLM), es crucial fortalecer las capacidades de computación de alto rendimiento e implementar políticas de gobernanza de datos que respeten los derechos humanos y promuevan el uso responsable y soberano de los datos, favoreciendo así la investigación, el desarrollo tecnológico y la cooperación regional (cf. Roadmap for Ethical Artificial Intelligence for Latin America and the Caribbean 2024–2025, 2024, pp. 6-7).

Dentro de este marco de acción, algunos países y organizaciones han asumido un papel protagónico en la implementación de la hoja de ruta. Chile ha liderado este esfuerzo, posicionándose como referente en políticas de IA desde la adopción de su estrategia nacional en 2021. Su papel en la organización del foro y en el impulso de esta agenda ha reforzado su liderazgo en América Latina. Por su parte, la CAF ha destacado que la IA debe ser un factor de competitividad, desarrollo e inclusión en la región, promoviendo la implementación de infraestructuras tecnológicas, marcos éticos y estrategias de talento para IA. Con la consolidación del Consejo Regional y la aplicación de la hoja de ruta, América Latina y el Caribe buscan no solo regular la IA en sus territorios, sino también influir en las discusiones globales sobre su desarrollo y uso ético (cf. UNESCO, 2024).

11.1.5 Colombia

El Marco Ético para la IA en Colombia (2021) reconoce que el país ya cuenta con regulaciones relevantes para la protección de datos y la ciberseguridad. Entre ellas, destacan la

Ley 1581 de 2012, la Ley de Habeas Data y diversas normativas en ciberseguridad. Sin embargo, estas normativas no fueron diseñadas específicamente para la IA, por lo que es necesario adaptarlas a los desafíos actuales de esta tecnología.

La Ley 1581 de 2012 establece el régimen general de protección de datos personales en Colombia, garantizando el derecho fundamental de las personas a conocer, actualizar y rectificar la información que se haya recogido sobre ellas en bases de datos o archivos de entidades públicas o privadas. Esta normativa impone principios como legalidad, finalidad, libertad, veracidad, seguridad, confidencialidad y acceso restringido en el tratamiento de datos personales. Su aplicación es clave en el desarrollo de la IA, ya que regula cómo los sistemas de IA pueden recopilar, almacenar y utilizar información personal. También impone obligaciones a las empresas y entidades que desarrollan estas tecnologías, asegurando que protejan la privacidad de los ciudadanos (cf. Congreso de Colombia, 2012).

Por su parte, la Ley de Habeas Data (Ley 1266 de 2008) regula el manejo de la información contenida en bases de datos de carácter financiero, crediticio, comercial y de servicios, estableciendo derechos y procedimientos para garantizar la protección de los datos personales en estos ámbitos. Mientras que la Ley 1581 regula el tratamiento de datos personales en general, la Ley de Habeas Data se centra en la información financiera y crediticia. En el contexto de la IA, esta ley es clave para evitar que algoritmos utilizados en el sector financiero tomen decisiones discriminatorias o poco transparentes sobre créditos y servicios financieros (cf. Congreso de Colombia, 2008).

El Marco Ético para la IA en Colombia (2021) resalta que, aunque en Colombia existen algunas leyes que pueden aplicarse a la IA, estas no han sido diseñadas específicamente para abordar sus desafíos. Por esta razón, se recomienda actualizar y mejorar las normativas existentes

para garantizar que la IA se use de manera ética y segura. Se sugiere que el gobierno implemente regulaciones más claras y adaptadas a los avances tecnológicos, asegurando que protejan los derechos de las personas sin limitar la innovación (Ministerio de Ciencia, Tecnología e Innovación, 2021, pp. 52-60).

Además, el documento establece cuatro principios clave para garantizar un desarrollo responsable de la IA en Colombia:

- **Transparencia:** Los sistemas de IA deben ser comprensibles para los ciudadanos, es decir, se debe poder explicar cómo toman decisiones y qué datos utilizan.
- **Responsabilidad:** Las empresas y entidades que desarrollan o usan IA deben asumir la responsabilidad por sus impactos y posibles errores.
- **Seguridad:** Se deben implementar medidas para evitar que la IA sea utilizada con fines dañinos o cause riesgos para las personas.
- **Respeto a los derechos humanos:** La IA no debe ser utilizada para discriminar ni afectar negativamente la dignidad, la privacidad o la libertad de las personas.

  (cf. Ministerio de Ciencia, Tecnología e Innovación, 2021, pp. 22-38).

El Ministerio de Ciencia, Tecnología e Innovación (2021) presentan recomendaciones concretas para la creación de nuevas regulaciones en IA. En lugar de imponer normas rígidas que podrían frenar la innovación, se sugiere diseñar un marco normativo flexible y adaptable que permita aprovechar los beneficios de la IA sin comprometer la seguridad ni los derechos de las personas. Además, se plantea la necesidad de contar con mecanismos de auditoría y certificación, es decir, procesos que permitan evaluar y verificar que los sistemas de IA sean confiables, justos y no generen sesgos o discriminación (cf. pp. 52-60).

11.2 Organismos internacionales y normas globales

11.2.1 ONU

Según Consejo de Derechos Humanos de las Naciones Unidas (2021), el uso de la IA debe regirse por principios fundamentales de derechos humanos como la legalidad, legitimidad, necesidad y proporcionalidad. La Alta Comisionada de las Naciones Unidas para los Derechos Humanos enfatiza que la protección efectiva del derecho a la privacidad y derechos conexos frente a la IA depende de marcos jurídicos, reglamentarios e institucionales sólidos establecidos por los Estados (cf. p. 12).

Las leyes de protección de datos deben adaptarse a los desafíos específicos que plantea la IA, como el uso de inferencias automatizadas o decisiones sin intervención humana. Se propone, por ejemplo, que las normativas reconozcan el derecho de las personas a obtener una explicación significativa y a oponerse a decisiones totalmente automatizadas (cf. Consejo de Derechos Humanos de las Naciones Unidas, 2021, p. 12).

Además, la Alta Comisionada considera que los sectores de alto riesgo, como la seguridad nacional, la justicia penal, la asistencia social o el empleo, deben estar regulados por marcos más estrictos, e incluso prohibir ciertos usos de la IA incompatibles con los derechos humanos, como los sistemas de puntuación social o algoritmos que clasifican a las personas de forma discriminatoria. También recomienda moratorias sobre tecnologías de alto riesgo, como el reconocimiento facial remoto en tiempo real, hasta que se demuestre que no vulneran derechos fundamentales (cf. Consejo de Derechos Humanos de las Naciones Unidas, 2021, p. 13). Recalca que la regulación debe ser intersectorial, con mecanismos de supervisión independientes y especializados, incluyendo autoridades de protección de datos y organismos de derechos humanos y consumidores (cf. Consejo de Derechos Humanos de las Naciones Unidas, 2021, p. 14).

Complementariamente, en la Asamblea General de las Naciones Unidas (2024) subraya la necesidad urgente de establecer sistemas de IA seguros, protegidos y fiables, entendidos como aquellos que respetan los derechos humanos, el derecho internacional, la privacidad, y que son éticos, inclusivos, centrados en las personas y orientados al desarrollo sostenible (cf., p. 3).

La ONU reconoce tanto el potencial transformador como los riesgos inherentes al uso de la IA. Señala que su utilización incorrecta o maliciosa, sin salvaguardias adecuadas, puede reforzar desigualdades estructurales, agrandar brechas digitales y vulnerar derechos humanos, incluido el derecho a la privacidad. Por ello, destaca la urgencia de alcanzar un consenso mundial sobre normas internacionales que sean eficaces, interoperables, inclusivas y adaptables, para evitar una gobernanza fragmentada (cf. Asamblea General de las Naciones Unidas, 2024, p. 3).

La resolución exhorta a los Estados a respetar y proteger los derechos humanos durante todo el ciclo de vida de los sistemas de IA y a abstenerse de usar tecnologías incompatibles con el derecho internacional o que representen riesgos indebidos para poblaciones vulnerables (cf. Asamblea General de las Naciones Unidas, 2024, p. 5). También alienta la elaboración de marcos regulatorios nacionales que promuevan la transparencia, la explicabilidad, la supervisión humana y la protección de datos personales, incluyendo evaluaciones de impacto ético y social (cf. Asamblea General de las Naciones Unidas, 2024, p. 6).

De igual forma, se destaca el papel de la cooperación internacional como mecanismo clave para cerrar las brechas digitales y garantizar la participación equitativa de los países en desarrollo en los foros regulatorios sobre IA, mediante asistencia técnica, financiera y transferencia tecnológica (cf. Asamblea General de las Naciones Unidas, 2024, pp. 4-5). Finalmente, la ONU reafirma que su sistema institucional debe liderar la construcción de un consenso global en gobernanza de la IA, en coherencia con la Carta de las Naciones Unidas, la Declaración Universal

de Derechos Humanos y la Agenda 2030 para el Desarrollo Sostenible (Asamblea General de las Naciones Unidas, 2024, p. 9).

11.2.2 UNESCO

Según la UNESCO (2022) reconoce que los sistemas de IA, aunque ofrecen oportunidades significativas para el desarrollo, también plantean riesgos éticos como la reproducción de sesgos, la exclusión, la vigilancia masiva y la afectación a derechos fundamentales. Por ello, la UNESCO promueve un enfoque regulador basado en valores como la dignidad humana, la inclusión, la equidad, la sostenibilidad ambiental y la transparencia, a lo largo de todo el ciclo de vida de estos sistemas (cf. pp. 17–20).

Uno de los pilares fundamentales de la recomendación es el principio de proporcionalidad e inocuidad, que exige que el uso de la IA sea adecuado al contexto, no excesivo y no infrinja los derechos humanos. Se prohíbe explícitamente el uso de la IA para sistemas de puntuación social o vigilancia masiva (cf. UNESCO, 2022, p. 20).

Además, la UNESCO (2022) promueve la creación de marcos de evaluación del impacto ético, recomienda a los Estados Miembros a evaluar los beneficios y riesgos de los sistemas de IA, especialmente en relación con los derechos humanos, el medio ambiente y los efectos en poblaciones vulnerables. Estas evaluaciones deben ser transparentes, participativas e inclusivas, y contemplar auditorías, trazabilidad y mecanismos de supervisión adecuados (cf. pp. 26–27).

La UNESCO (2022) señala la importancia del derecho a la privacidad y la protección de datos, e recomienda a los Estados a establecer marcos de gobernanza adecuados y a garantizar evaluaciones de impacto sobre la privacidad en todas las etapas del ciclo de vida de los sistemas de IA. También la recomendación enfatiza la necesidad de transparencia, explicabilidad,

supervisión humana y rendición de cuentas como condiciones esenciales para garantizar la fiabilidad y el uso ético de la IA. Se exige que las decisiones automatizadas puedan ser comprendidas, cuestionadas y revisadas por humanos, especialmente cuando afectan derechos fundamentales (cf. pp. 22–23).

11.2.3 OECD

La OECD (2019) emitió la primera norma intergubernamental sobre IA, este instrumento tiene como propósito fomentar la innovación responsable y la confianza en sistemas de IA confiables, al tiempo que se garantiza el respeto por los derechos humanos y los valores democráticos (cf. p. 3).

La recomendación se estructura en dos secciones fundamentales. La primera expone cinco principios de alto nivel para una gestión responsable de la IA confiable, aplicables a todos los actores involucrados:

- **Crecimiento inclusivo, desarrollo sostenible y bienestar:** Se promueve que la IA fortalezca capacidades humanas, reduzca desigualdades y proteja el medio ambiente (cf., OECD, 2019, p. 8).
- **Respeto del Estado de derecho, los derechos humanos y los valores democráticos:** Se exige que los sistemas de IA respeten principios como la equidad, la privacidad, la autonomía individual y la no discriminación a lo largo de su ciclo de vida. También se incluyen mecanismos para enfrentar el uso indebido o malintencionado de estos sistemas (cf., OECD, 2019, p. 8).
- **Transparencia y explicabilidad:** Los actores de la IA deben proporcionar información clara sobre el funcionamiento y decisiones de los sistemas, permitiendo que los afectados

comprendan y, cuando sea necesario, impugnen los resultados generados por IA (cf., OECD, 2019, p. 8).

- **Robustez, seguridad y protección:** Los sistemas de IA deben operar de forma segura en condiciones normales y adversas, con mecanismos que permitan su desactivación, reparación o retiro cuando representen riesgos (cf., OECD, 2019, p. 9).
- **Responsabilidad:** Los actores deben ser responsables del funcionamiento adecuado de los sistemas, asegurando trazabilidad, gestión de riesgos, y conducta empresarial responsable en todo el ciclo de vida (cf., OECD, 2019, p. 9).

La segunda sección contiene cinco recomendaciones para los países adherentes, enfocadas en políticas nacionales y cooperación internacional:

- **Invertir en investigación y desarrollo de IA:** Se promueve el financiamiento público y privado en investigación interdisciplinaria y herramientas de código abierto que fomenten una IA confiable y sin sesgos (cf., OECD, 2019, p. 9).
- **Fomento de un ecosistema inclusivo que facilite la IA:** Se debe construir una infraestructura interoperable para compartir datos, tecnología y conocimiento de IA de forma segura y ética (cf., OECD, 2019, p. 10).
- **Desarrollo de un entorno de gobernanza y políticas interoperable y propicio para la IA:** Los gobiernos deben adaptar sus marcos regulatorios para permitir el desarrollo seguro de la IA, incluyendo entornos de prueba y normativas basadas en resultados (cf., OECD, 2019, p. 10).
- **Desarrollo de la capacidad humana y preparación para la transformación del mercado laboral:** Se busca preparar a la sociedad para los cambios del mercado laboral

derivados de la IA, mediante formación continua, protección social y promoción de empleos de calidad (cf., OECD, 2019, p. 10).

- **Cooperación internacional para una IA fiable:** Se alienta a los Estados a trabajar conjuntamente en foros multilaterales para compartir conocimientos, desarrollar estándares técnicos y promover la adopción de IA confiable (cf., OECD, 2019, p. 10).

Finalmente, la OECD instruye a su Comité de Política Digital a supervisar la implementación, actualizar la recomendación según avances tecnológicos (como la IA generativa), y proporcionar orientación práctica a los Estados (cf., OECD, 2019, p. 11).

11.3 Desafíos y debates actuales

Uno de los problemas que Crawford (2021) destaca es el daño ambiental que genera la IA, explica que la fabricación de hardware y el entrenamiento de modelos de IA requieren la extracción de minerales, el uso masivo de energía y grandes cantidades de agua. La expansión de la IA está ligada a la crisis climática, lo que plantea la necesidad de desarrollar tecnologías más sostenibles y responsables con el planeta (cf. pp. 23-52).

Crawford (2021) expone que la IA no es totalmente autónoma, sino que depende del trabajo de muchas personas que realizan tareas esenciales, pero poco reconocidas. Por ejemplo, trabajadores en países en desarrollo etiquetan datos para entrenar algoritmos o moderan contenido en redes sociales. Este tema genera discusión sobre la necesidad de mejores condiciones laborales y regulaciones que protejan a quienes sostienen la economía de datos (cf. pp. 53-88).

Crawford (2021) explica cómo los algoritmos pueden reforzar discriminación racial, de género y de clase. Discute cómo los sistemas de IA, al estar entrenados con datos históricos que

reflejan desigualdades sociales, pueden perpetuar patrones de exclusión. Este problema plantea la necesidad urgente de crear modelos más justos y éticos (cf. pp. 123-150).

Crawford (2021) analiza cómo la IA se usa para vigilar a la población. Gobiernos y empresas recopilan grandes cantidades de datos personales, lo que genera preocupaciones sobre la privacidad y el control social. Estos sistemas han llevado a debates sobre los límites entre seguridad y derechos individuales. La expansión de la vigilancia basada en IA plantea preguntas sobre el equilibrio entre la protección de la seguridad pública y la preservación de los derechos individuales (cf. pp. 181-210).

Crawford (2021) habla sobre el dominio de la IA por parte de grandes empresas tecnológicas como Google, Microsoft y Amazon. Estas compañías no solo lideran la investigación en IA, sino que también influyen en las leyes y regulaciones que afectan su uso. Este control sobre la tecnología genera preocupación sobre cómo democratizar su acceso y garantizar que beneficie a más personas en lugar de concentrar el poder en unas pocas empresas (cf. pp. 181-228).

11.4 Principios éticos en la regulación de la IA

Müller (2020) habla de lo justo que debe ser todo y de no discriminar; analiza por qué la IA a veces se equivoca y cómo puede empeorar las cosas si no se diseña pensando en todos. Si la IA decide quién trabaja, quién recibe justicia o quién se atiende en un hospital, puede haber problemas. Es vital evitar estos riesgos y asegurar que la IA trate a todos por igual, sin importar las características de las personas (cf. pp. 8-9).

Müller (2020) pregunta algo que da qué pensar: ¿a quién culpar si la IA comete un error o lastima a alguien? Como estas máquinas piensan solas, es difícil saber quién es el responsable, lo que ha creado discusiones sobre cómo controlar la IA con leyes nuevas. Müller insiste en que

quienes crean y usan la IA deben responder por sus actos y buscar formas de arreglar errores y cuidar a quienes puedan salir perjudicados (cf. pp. 11-12).

11.5 Casos de estudio

Zuboff (2019) revela que en las páginas web se extrae información, donde la IA digiere esta información gigante, afinando trucos de venta. Las empresas no solo predicen los comportamientos de los usuarios, sino que diseñan experiencias que los mantienen enganchados, generando aún más datos y favoreciendo una retroalimentación constante entre el usuario y la plataforma. Esto implica que las plataformas pueden, en esencia, “hacer bailar” a los usuarios, alentando acciones específicas a través de recomendaciones personalizadas, publicidad dirigida y contenido viral. Esto muestra el poder que tienen los algoritmos sobre las decisiones y comportamientos humanos, un poder que debe ser regulado para evitar abusos y manipulación emocional (cf. Zuboff ,2019, pp. 131-150).

Zuboff (2019) explica cómo la IA es utilizada para automatizar la toma de decisiones en áreas críticas, como el empleo, el crédito y la justicia penal. Aunque estos sistemas prometen eficiencia, también generan riesgos de sesgo y discriminación, ya que se basan en datos históricos que pueden estar plagados de prejuicios. El uso de la IA en estos contextos subraya la necesidad de que los gobiernos implementen regulaciones que aseguren que estos modelos sean auditables, transparentes y justos, de modo que no perpetúen desigualdades existentes (cf. pp. 279-296).

Zuboff (2019) describe como un “golpe desde arriba”. Las empresas tecnológicas tienen tanto poder que pueden moldear las reglas a su favor, manipulando las políticas gubernamentales para que se alineen con sus intereses. Esto resalta la importancia de que las políticas de IA sean independientes de los intereses corporativos y que se protejan los derechos de los usuarios frente

a los poderes de las grandes plataformas. La regulación debe asegurar que las tecnologías emergentes, como la IA, no sean usadas para manipular a los ciudadanos, sino para promover un entorno más justo y transparente (cf. pp. 309-326).

12. Ética en el diseño y desarrollo de sistemas de IA

Es bien sabido que cada día la IA sigue transformando diversos sectores, ofreciendo soluciones innovadoras y eficientes. Pero, su avance tecnológico y su potencial disruptivo plantean importantes desafíos éticos que se deben abordar desde las etapas iniciales y desarrollo de los sistemas de IA. A continuación, se explora los principios éticos fundamentales que guían su desarrollo, con especial atención al enfoque *Ethics by Design*, que emerge como respuesta estructural proponiendo la integración proactiva de principios éticos en el ADN de los sistemas de IA y su importancia en la creación de sistemas responsables que respeten los valores humanos y promuevan el bienestar social.

Integrar criterios éticos en la planificación y la creación de sistemas de IA se ha convertido en algo imprescindible para asegurar que estas herramientas ayuden a la sociedad sin acarrear consecuencias negativas graves. Los estudios actuales muestran que las estrategias más exitosas son las que incluyen la ética desde las primeras fases de creación, usando modelos de trabajo variados y métodos de evaluación continua. Los criterios esenciales como la ecuanimidad algorítmica, la claridad, la intimidad, la seguridad y la independencia responsable deben orientar todo el proceso de los sistemas de IA.

12.1 Contexto y evolución de la Ética en los sistemas de IA

La IA ha emergido como una tecnología transformadora con potencial para renovar múltiples aspectos de la sociedad actual. Sin embargo, esta capacidad viene acompañada de profundos desafíos éticos que se deben abordar de manera sistemática y proactiva. Las implicaciones éticas de la IA no son meramente teóricas, sino que impactan directamente en como estas tecnologías moldean las relaciones sociales, los procesos de toma de decisiones y los derechos fundamentales de las personas.

La evolución del discurso ético en torno a la IA ha transitado desde enfoques reactivos, que abordan problemas éticos una vez manifestados, hacía paradigmas preventivos que buscan anticipar y mitigar dilemas éticos desde las fases iniciales de conceptualización y diseño. Esta transición refleja la comprensión de que los sistemas de IA no son éticamente neutros, sino que incorporan valores, sesgos y prioridades —explícitas o implícitas— de quienes los diseñan y desarrollan.

El desarrollo ético de los sistemas de IA requiere de un enfoque integral que considere no solo aspectos técnicos sino también dimensiones sociales, culturales y legales. Los principios éticos no deben considerarse como restricciones al desarrollo tecnológico, sino como guías para crear sistemas robustos, confiables y beneficiosos para la humanidad.

12.2 Principios éticos Fundamentales en la IA

12.2.1 Justicia y Equidad Algorítmica

La imparcialidad en los algoritmos es clave para una evolución ética de la IA, buscando que operen de forma justa y sin exacerbar las diferencias ya presentes. Esto implica examinar a

fondo los datos de entrenamiento, ya que pueden reflejar sesgos arraigados que, sin detección ni corrección, se repetirán y agravarán en los sistemas de IA.

Su progreso debe integrar métodos para valorar la existencia de sesgos en cada fase de la IA, desde la elección y adecuación de datos hasta su implementación y supervisión constante. Esto involucra tanto análisis técnicos como la comprensión de cómo las elecciones algorítmicas impactan a diversos grupos sociales.

12.2.2 Transparencia y Aplicabilidad

La transparencia y aplicabilidad son principios interrelacionados que abordan la necesidad de que los sistemas de IA sean comprensibles y auditables tanto por expertos como por usuarios y personas afectadas por sus decisiones. La implementación de estos principios implica superar la percepción de los sistemas de IA como “cajas negras” inexplicables.

La transparencia abarca aspectos como la documentación clara sobre los objetivos del sistema, las fuentes de datos utilizadas, las metodologías de desarrollo y los criterios de evaluación implementados. Esto facilita la rendición de cuentas y permite a las partes interesadas evaluar la adecuación y legitimidad del sistema para contextos específicos.

La aplicabilidad, por su parte, se refiere a la capacidad del sistema para proporcionar justificaciones comprensibles sobre sus procesos de toma de decisiones. Esto es particularmente crucial en los sistemas que afectan derechos o intereses significativos de las personas, como aquellos utilizados en ámbitos de salud, justicia o acceso a recursos públicos.

Es importante señalar que el nivel adecuado de transparencia y aplicabilidad debe calibrarse según el contexto y propósito específico del sistema, reconociendo potenciales tensiones

con otros principios como la privacidad o la seguridad. No obstante, incluso en sistemas complejos como DNN, se deben incorporar herramientas y metodologías que faciliten su interpretación.

12.2.3 Privacidad y Protección de Datos

En el desarrollo ético de los sistemas de IA, la privacidad y la protección de datos representa preocupaciones fundamentales que se deben abordar, igualmente, de forma sistemática. Los sistemas de IA frecuentemente procesan grandes volúmenes significativos de datos personales, lo que implica riesgos sustanciales para la privacidad individual y colectiva si no se implementan protecciones adecuadas.

La protección de la privacidad debe considerarse durante todo el ciclo de vida de los sistemas de IA, desde la recolección de los datos hasta su implementación y posterior evolución. Esto requiere adoptar enfoques como la privacidad por diseño y por defecto, que incorporan consideraciones de privacidad desde las etapas iniciales de conceptualización.

Los marcos adecuados de protección de datos deben establecer límites claros sobre qué datos se puede recopilar, cómo se deben almacenar, quién puede acceder a ellos y la finalidad con los que serán utilizados.

12.2.4 Seguridad y Fiabilidad

La seguridad constituye un principio ético fundamental en el desarrollo de sistemas de IA, entendida como la obligación de crear tecnologías que no generen daños, peligros, riesgos o amenazas como consecuencia de su uso. Este principio abarca tanto la integridad física como el bienestar psicológico y social de las personas.

Crear sistemas de IA seguros requiere usar métodos estrictos para valorar los riesgos. Se deben buscar puntos débiles a nivel técnico fallos en el sistema, vulnerabilidades o ataques (y a nivel social efectos inesperados o usos indebidos). Esta valoración debe ser constante y ajustarse a nuevas amenazas y formas de uso que aparezcan tras la primera puesta en marcha del sistema.

La fiabilidad complementa el principio de seguridad, refiriéndose a la consistencia y previsibilidad con que los sistemas de IA funcionan según lo esperado bajo diversas condiciones. Para garantizar la fiabilidad, es esencial implementar pruebas exhaustivas que evalúen el rendimiento del sistema en escenarios diversos, incluidos casos extremos y situaciones imprevistas.

12.2.5 Autonomía y Responsabilidad

La autonomía y responsabilidad constituyen principios complementarios que abordan la tensión entre la capacidad de los sistemas de IA para operar independientemente y la necesidad de mantener la rendición de cuentas humana sobre sus acciones y consecuencias. Estos principios cobran especial relevancia a medida que los sistemas de IA adquieren mayores capacidades para tomar decisiones sin intervención humana directa.

En este contexto, la autonomía tiene una doble dimensión: por un lado, se refiere a la capacidad técnica de los sistemas para funcionar con cierto grado de independencia; por otro lado, se refiere al respeto por la autonomía humana, que implica que las personas mantengan control significativo sobre sus decisiones que afectan sus vidas y no sean manipuladas o subordinadas a procesos automáticos.

El principio de responsabilidad establece que siempre debe ser posible atribuir responsabilidad ética y jurídica por las acciones y consecuencias de los sistemas de IA a personas

físicas o entidades jurídicas específicas. Esto requiere mecanismos claros de supervisión humana, especialmente en sistemas con alto impacto potencial, así como estructuras de gobernanza que definan roles y responsabilidades específicas.

12.3 *Ethics by Design for AI* —EbD-AI—

En la era de los sistemas de IA y la digitalización global, el concepto de *Ethics by Design* (Ética por Diseño) surge como un paradigma esencial para garantizar que el desarrollo tecnológico priorice el bienestar humano y social. Este enfoque propone integrar consideraciones éticas desde las primeras etapas de diseño de sistemas tecnológicos, en lugar de abordarlas como una revisión posterior. Como señala la Comisión Europea (2021), "la ética debe ser un componente intrínseco del ciclo de vida tecnológico, no un parche aplicado ante crisis" (p. 5).

Para Brey y Dainow (2023), "*Ethics by Design* es un enfoque que busca incluir sistemáticamente y de manera integral consideraciones éticas en el proceso de diseño y desarrollo de nuevos sistemas y dispositivos tecnológicos. Aunque este enfoque se puede aplicar a cualquier tecnología, históricamente se ha centrado en el diseño de sistemas de IA". Este paradigma sosteniente que la ética debe integrarse de manera transversal y estructural en cada fase del ciclo de vida del sistema —desde la definición de los objetivos y la selección de los datos, hasta el diseño algorítmico, la implementación y el monitoreo continuo—.

Ethics by Design se enraíza en movimientos previos como *Privacy by Design*, conceptualizado por Cavoukian (2011), cuyos siete principios buscan proteger la privacidad mediante diseño técnico proactivo. Sin embargo, su alcance se expande para abarcar valores como transparencia, equidad y responsabilidad. Floridi (2018) argumenta que en la "cuarta revolución" digital, la tecnología redefine nuestra ontología, exigiendo marcos éticos que trasciendan lo

humano y lo artificial, es decir, que la realidad ya no se limita a lo físico o biológico, sino que se expande a un universo informacional (infosfera) donde humanos, máquinas y datos coexisten como entidades interdependientes.

Un pilar clave es el *Value Sensitive Design* (Diseño Sensible a Valores), que, según Jacobs (2021), "integra valores éticos en la arquitectura técnica, asegurando que las decisiones de diseño reflejen prioridades sociales" (párr. 3). Este enfoque se complementa con el marco ART (*Accountability, Responsibility, Transparency*) de Dignum (2019), que enfatiza la rendición de cuentas, responsabilidad y transparencia en sistemas autónomos.

Por otro lado, *Ethics by Design* es un enfoque desarrollado en colaboración con los proyectos SHERPA y SIENNA financiados por la UE que propone un marco para asegurar que las cuestiones éticas se aborden a lo largo de todo el proceso de desarrollo de la IA. El proyecto SHERPA, se dedica a analizar cómo la IA y el análisis de grandes volúmenes de datos impactan en la ética y los derechos humanos. SHERPA trabaja en diálogo con múltiples actores para desarrollar nuevas formas de entender y abordar estos desafíos, con el fin de encontrar soluciones sostenibles y beneficiosas tanto para los innovadores como para la sociedad en general. En este sentido, SHERPA promueve explícitamente la aplicación de los principios de *Ethics by Design*, asegurando que las tecnologías emergentes cumplan con estándares éticos desde su concepción y durante todo su ciclo de vida. Además, el proyecto ha contribuido a la elaboración de directrices y recomendaciones para el desarrollo ético y responsable de la IA, alineándose con marcos regulatorios como Horizonte Europa *(Project SHERPA, 2021; European Commission, s. f.*).

Por su parte, SIENNA aborda los aspectos éticos, legales y sociales de la IA, el *big data* y otras tecnologías emergentes, con el objetivo de proporcionar un marco normativo y recomendaciones para que estas tecnologías respeten los derechos fundamentales y fomenten la

confianza pública. SIENNA enfatiza la importancia de la gobernanza responsable y la participación de diversos actores sociales para implementar la ética de manera integral en el diseño y uso de tecnologías avanzadas, lo que coincide con los principios de *Ethics by Design* que buscan integrar la ética como un componente esencial y no como una consideración secundaria (*European Commission*, 2020).

El proceso para aplicar *Ethics by Design* se compone de varias etapas interrelacionadas que permiten integrar principios éticos desde la concepción hasta la implementación de sistemas tecnológicos. Estas fases, que incluyen evaluación, instanciación, mapeo, aplicación e implementación, buscan garantizar que los valores éticos se traduzcan en requisitos técnicos concretos y acciones prácticas durante el desarrollo.

El primer paso, la **evaluación**, consiste en identificar y analizar los valores éticos relevantes para el proyecto, así como los posibles riesgos y oportunidades asociados. Esta fase implica un estudio profundo del contexto social, legal y técnico, para comprender qué aspectos éticos deben priorizarse. Según Brey y Dainow (2023), esta evaluación inicial es fundamental para orientar el diseño hacia principios como la equidad, la privacidad y la transparencia.

A continuación, en la fase de **instanciación**, los valores abstractos detectados se traducen en requisitos técnicos específicos. Por ejemplo, si la privacidad es un valor clave, se definirán mecanismos concretos para proteger datos sensibles mediante técnicas de anonimización o cifrado. Este paso es crucial para transformar conceptos éticos en parámetros medibles y verificables (*European Commission*, 2021).

El **mapeo** es la etapa donde se asignan responsabilidades y se diseñan las acciones concretas para cumplir con los requisitos éticos. Esto implica distribuir tareas entre los equipos de

desarrollo, definir protocolos de supervisión y establecer indicadores para monitorear el cumplimiento ético a lo largo del ciclo de vida del sistema (Brey y Dainow, 2023).

La **aplicación** se refiere a la integración efectiva de estas acciones y requisitos en los procesos de desarrollo, utilizando metodologías ágiles o tradicionales según corresponda. Aquí, las revisiones éticas se convierten en parte de los *sprints* o ciclos de trabajo, permitiendo ajustes continuos basados en retroalimentación y pruebas (*European Commission*, 2021).

Finalmente, la **implementación** implica la puesta en marcha del sistema con mecanismos de monitoreo y evaluación continua para asegurar que los principios éticos se mantengan vigentes en operación real. Se utilizan *dashboards*, auditorías y reportes para detectar desviaciones y corregirlas oportunamente, garantizando así la sostenibilidad ética del sistema (Brey y Dainow, 2023).

Ahora bien, los desarrolladores y diseñadores asumen una responsabilidad activa para anticipar posibles impactos éticos negativos, tales como sesgos, falta de transparencia o riesgos para la privacidad, y para implementar mecanismos técnicos y organizativos que mitiguen esos riesgos. Por ejemplo, en la capa de datos, se aplican técnicas de anonimización —métodos utilizados para modificar datos personales con el fin de proteger la privacidad y corregir sesgos estructurales—; en el diseño algorítmico, se incorporan funciones de pérdida que penalizan decisiones injustas o discriminatorias; y en la interfaz, se desarrollan sistemas explicativos que faciliten la comprensión y la rendición de cuentas hacía los usuarios.

El proyecto *TechEthos*, financiado por la UE, ejemplifica la aplicación práctica de *Ethics by Design*. Según ALLEA (2023), este proyecto generó directrices éticas para tecnologías emergentes como la ingeniería climática y las neurotecnologías, destacando la necesidad de "equilibrar innovación con salvaguardias éticas" (párr. 4).

Es cierto que muchas de las consecuencias del desarrollo de una tecnología dependen parcial o totalmente de cómo y en qué contexto se usa, sin embargo, el diseño también importa; las decisiones de diseño algunas veces pueden generar consecuencias particulares en una gran diversidad y contextos de uso. Por ejemplo, los algoritmos de recomendación en redes sociales pueden ser diseñados para maximizar el tiempo de usuario o para promover contenido diverso; si se prioriza el tiempo de usuario, podría llevar a la creación de "burbujas de filtro"[21] que refuerzan en los usuarios esos sesgos existentes y reducen la exposición a perspectivas diversas, lo que puede tener impactos negativos en la cohesión social y la democracia. De igual forma, un sistema de diagnóstico médico puede ser entrenado con datos que reflejen sesgos raciales o de género; si el sistema perpetúa estos sesgos, podría dar lugar a diagnósticos inexactos o tratamientos desiguales para ciertos grupos, lo que tendría graves consecuencias éticas y de salud pública.

Ethics by Design no es una opción, sino una responsabilidad colectiva. Al integrar principios éticos en cada línea de código, las sociedades pueden evitar escenarios distópicos donde la tecnología domine en lugar de servir. Como concluye Dignum (2019), "la IA responsable no es un lujo, sino un requisito para la sostenibilidad digital" (p. 143). El futuro exige una alianza entre ingenieros, filósofos y legisladores para construir tecnologías que reflejen lo mejor de la humanidad, no sus limitaciones.

21 Estados de aislamiento intelectual generados por algoritmos que personalizan la información mostrada al usuario según su comportamiento previo, limitando la exposición a perspectivas diversas y reforzando sesgos personales (Pariser, 2011).

13. IA en la toma de decisiones y gobernanza algorítmica

En los núcleos centrales de la interacción social, la incorporación de sistemas de IA en los procedimientos decisorios de las instituciones está generando un cambio significativo en las fuerzas de poder, la rendición de cuentas y la validez. Este cambio va más allá de una simple mejora técnica; implica una reorganización sustancial del poder de decisión entre personas, herramientas tecnológicas y organizaciones. Frank Pasquale (2015), en su obra influyente (*The Black Box Society: The Secret Algorithms That Control Money and Information*), alerta sobre los efectos en la democracia de encomendar decisiones cruciales al control poco transparente de los algoritmos, sobre todo cuando afectan derechos fundamentales o la distribución de recursos sociales limitados.

La integración de estas innovaciones en áreas como la valoración crediticia, la asignación de fondos públicos o las elecciones de personal no solo pone a prueba su robustez técnica, sino sobre todo su compromiso con valores democráticos como la claridad, la responsabilidad y la participación de la gente en la definición de las normas que deben regir las decisiones que nos afectan a todos. Cuando estos sistemas actúan como "cajas negras" difíciles de entender, la habilidad de los afectados para entender, cuestionar o rebatir las decisiones que influyen en sus vidas disminuye, dando lugar a lo que Mireille Hildebrandt (2020) denominó un "déficit democrático algorítmico" en su libro Tecnologías Inteligentes y el Fin de la Ley.

Una tensión entre eficiencia técnica y legitimidad democrática es particularmente clara en la esfera pública, donde el uso de sistemas algorítmicos para el control de programas sociales está remodelando fundamentalmente la relación entre ciudadanos e instituciones. En *Algorithms in Practice*, Christin (2017) examina cómo la adopción de herramientas predictivas en servicios

sociales remodela las ideas de necesidad, merecimiento y responsabilidad que subyacen a las políticas de bienestar.

Codificar algorítmicamente factores cuantificables y relaciones estadísticas como criterios evaluativos los privilegia sobre el contexto o la narración personal, reconstituyendo ambos, lo que es evidencia relevante para la toma de decisiones públicas. Esta tendencia hacia la cuantificación y la metodologización puede desplazar o eliminar ciertos conocimientos y formas de experiencia, especialmente relacionados con las vidas de grupos históricamente subordinados, como Broussard postula acertadamente en IA.

La creciente tendencia a confiar la potestad de decidir a sistemas basados en algoritmos está transformando el panorama de la responsabilidad profesional en ámbitos tradicionalmente caracterizados por una elevada independencia social y un criterio experto. En aquellos entornos institucionales donde las sugerencias algorítmicas coexisten con las evaluaciones humanas, se manifiestan patrones intrincados de aceptación, rechazo y diálogo que redefinen las fronteras de la autoridad del saber. Investigaciones empíricas sobre sistemas algorítmicos en las administraciones públicas, realizadas por Veale, Van Kleek y Binns (2018), señalan tendencias de "aceptación ciega" de las recomendaciones informáticas, evidenciando esta sumisión automática en contextos donde la institución se muestra reacia al riesgo o donde las presiones internas intensifican las altas cargas de trabajo y las limitaciones de tiempo.

Estas metodologías, denominadas por Citron como "automatización del criterio administrativo", pueden infringir requisitos fundamentales para mantener espacios de debate y contextualización necesarios al aplicar normas generales a casos concretos. La asombrosa habilidad de estos algoritmos para manejar enormes volúmenes de información y generar

resultados predictivos con una precisión aparente puede generar una sensación engañosa de perfección, lo cual desincentiva una supervisión humana realista.

13.1 Sistemas Económicos: La Evolución de los Bienes Públicos y el Papel de los Fallos del Mercado

El reconocimiento de los problemas fundamentales presentados por el continuo algoritmo de los procesos de toma de decisiones ha catalizado la aparición de marcos regulatorios, mecanismos institucionales, principios y estructuras de gobernanza que buscan regular dichos sistemas tecnológicos. Desde iniciativas sectoriales como los esfuerzos de autorregulación en torno a algoritmos hasta piezas legislativas de alto perfil como la propuesta de Regulación de IA de la Unión Europea, este campo emergente de la "gobernanza algorítmica" está moldeando de manera heterogénea y rápida un paisaje regulatorio.

En su influyente texto "Regulación Algorítmica", Yeung y Lodge proponen una valiosa clasificación que busca ordenar las diferentes maneras de regular, priorizando ya sean ideas generales, pasos procesales concretos o resultados que se puedan comprobar. Los modelos que se basan en ideas (como las pautas éticas para una IA fiable que publicó el Grupo de Expertos de Alto Nivel de la Comisión Europea) resaltan valores clave como la claridad, la justicia, la solidez técnica, la seguridad y el respeto por la libertad de las personas. Estos ofrecen un esquema regulatorio amplio, pero dejan mucho espacio para la interpretación sobre la forma en que se llevarán a cabo en la práctica.

Esta flexibilidad sin duda puede ayudar a adaptar las estructuras normativas a situaciones cambiantes y a tecnologías en constante evolución, pero por sí sola no asegura que las protecciones

sean efectivas, a menos que se complemente con métodos eficaces para comprobar que se cumplen las reglas y hacer que se respeten las normas.

En la elección, los regímenes procedimentales de regulación, como el creado por el Artículo 22 del Reglamento General de Protección de Datos de la Unión Europea, especifican requisitos (evaluaciones de impacto obligatorias, derechos de explicación o participación humana sustancial) para decisiones automatizadas con consecuencias legales significativas. Esta estrategia, que Edwards y Veale estudiaron de cerca en "Esclavo de Algoritmo", crea protecciones procedimentales diseñadas para ayudar a garantizar remedios efectivos para aquellos afectados por decisiones algorítmicas potencialmente dañinas.

Sin embargo, los estudios empíricos sobre cómo se utilizan estos mecanismos en la práctica, como el trabajo de Ausloos, Mahieu y Veale presentado en "*Getting Data Subject Rights Right*", muestran que hay grandes desconexiones entre los derechos que son reconocidos oficialmente y su cumplimiento efectivo, especialmente para personas en situaciones vulnerables o sin medios técnicos y legales adecuados.

En contraste, los enfoques orientados a resultados especifican umbrales mínimos de rendimiento verificables en dimensiones definidas (por ejemplo, precisión, equidad distributiva o robustez) independientemente de los procedimientos internos específicos que producen esos resultados. Propuestos por Kleinberg, Ludwig y Mullainathan en su artículo "Discriminación en la Era de los Algoritmos", este enfoque puede permitir la evaluación objetiva de si un conjunto de requisitos regulatorios está satisfecho, pero plantea desafíos metodológicos complejos sobre la operacionalización de conceptos normativamente cargados como la equidad o la no discriminación en métricas cuantificables y universalmente aplicables (Cox y Surace, 2022).

Una dimensión clave de la gobernanza algorítmica, más allá de enfoques regulatorios más explícitos, es la incorporación de una diversidad de perspectivas en el diseño, implementación y evaluación de dichas tecnologías a través de procesos de participación social. Costanza-Chock, en "Justicia del Diseño", argumenta de manera convincente que la participación sustantiva de los afectados por sistemas algorítmicos en su diseño no es solo otra obligación ética, sino una condición epistémica importante para descubrir y abordar impactos negativos que serán invisibles desde puntos de vista privilegiados.

Esta visión está respaldada por evidencia, como el trabajo realizado por Buolamwini y Gebru sobre el sesgo racial en el rendimiento de los sistemas de reconocimiento facial, sugiriendo que cuando los equipos técnicos carecen de diversidad demográfica, se crean puntos ciegos generalizados en la tarea de abordar riesgos discriminatorios.

Integrar diferentes perspectivas en el desarrollo y la puesta en marcha de algoritmos facilita un análisis más completo de los temas importantes. Además, posibilita una mejor comprensión de cómo estas tecnologías podrían impactar a la sociedad, sobre todo a las comunidades que han sido tradicionalmente excluidas. Los enfoques colaborativos también cumplen un papel crucial al validar estas herramientas, ya que ceder la facultad de decidir a sistemas automáticos a veces se interpreta como un menoscabo de principios democráticos básicos.

Mulgan, en "*Big Mind*", argumenta que la legitimidad social de los sistemas algorítmicos en dominios de alto impacto dependerá críticamente de la presencia de espacios deliberativos en los que diferentes actores sociales puedan negociar colectivamente los términos de esta delegación tecnológica.

Esta visión está alineada con una tradición de democracia deliberativa que busca legitimidad procedimental en procesos inclusivos de formación de voluntad colectiva

proporcionando así una conexión entre la gobernanza algorítmica y esta dimensión de la tradición democrática. Sin embargo, la manifestación real de tales ideales participativos debe lidiar con desafíos considerables asociados con las asimetrías de poder y conocimiento técnico y la opacidad que acompaña a muchos sistemas de aprendizaje automático.

Las posibilidades de estas tecnologías pueden tener complejidades técnicas que forman barreras a la participación que reflejan desigualdades existentes, centralizando la capacidad de influir en aquellos actores que actualmente tienen acceso privilegiado a los recursos cognitivos, técnicos y organizacionales que permitirían una participación significativa.

13.2 Perspectivas futuras

El desarrollo de sistemas de IA para informar y automatizar la toma de decisiones humanas está en un punto de inflexión donde ahora es más importante que nunca que todos los interesados reflexionen colectivamente sobre qué trayectorias tecnológicas en este espacio deseamos todos, individual y colectivamente, así como qué valores deberían fundamentar estas trayectorias. Las decisiones técnicas, institucionales y regulatorias tomadas durante este período formativo darán forma a las características de la infraestructura sociotécnica que estructurará muchos aspectos clave de la vida social durante décadas.

Por lo tanto, la industria y el gobierno deben abandonar enfoques tecno-deterministas que sostienen la evolución algorítmica basada como una trayectoria automática y necesaria, y en cambio adoptar perspectivas que entiendan el diseño e implementación tecnológica como elecciones inherentemente políticas.

Tal como la estudiosa del derecho Sheila Jasanoff explica convincentemente en su libro “La Ética de la Invención”, las tecnologías jamás son imparciales en cuanto a los valores. Más

bien, plasman e implementan nociones específicas acerca de vínculos sociales deseables, maneras legítimas de saber, y estándares para adjudicar recursos y chances. Entendido así, comprendemos que los sistemas algorítmicos no son solo herramientas técnicas que hacen posible alcanzar metas ya fijadas. También influyen al definir las pautas bajo las cuales moldeamos nuestra realidad, así como los caminos que deberíamos seguir para lograrlo.

Al analizar la faceta política intrínseca a la tecnología basada en algoritmos, esta perspectiva abre puertas para plantearnos preguntas esenciales. Cuestiona quién tiene el poder de decidir el rumbo de su desarrollo, qué ideas éticas se incorporan en su diseño, y qué fines se favorecen o se olvidan al usarla a gran escala. Zuboff, en su libro "La Era del Capitalismo de la Vigilancia", nos advierte sobre dejar decisiones tan importantes, basadas en elecciones sociales, solo en manos de las metas comerciales de aumentar las ganancias o la eficacia informática, sin considerar cómo esto afecta valores democráticos clave como la justicia, la libertad y el respeto humano. La concentración de capacidad técnica para los dispositivos de tales sistemas avanzados de IA entre un conjunto limitado de corporaciones tecnológicas pesa mucho en la gobernanza democrática de dichas tecnologías, particularmente cuando estas entidades operan con lógicas transnacionales que alteran las arquitecturas regulatorias territorialmente formateadas de los estados nacionales.

Ante tales desafíos, necesitamos establecer enfoques de gobernanza multinivel que incorporen reglas vinculantes, incentivos económicos, cambios organizacionales y procesos deliberativos que incluyan a todos. El registro de regulación en áreas como la seguridad alimentaria, los productos farmacéuticos o la aviación civil sugieren que es vital poner en marcha una institución o instituciones independientes con la experiencia técnica necesaria para evaluar y verificar el riesgo y para imponer sanciones que disuadan el comportamiento en casos

de violaciones graves.

Al analizar el aspecto político que reside en la tecnología algorítmica, notamos que su cualidad inherentemente difusa, a menudo opaca, y su evolución constante, generan desafíos particulares para su gestión, exigiendo adaptaciones significativas en las estructuras institucionales. Estrategias innovadoras, como la instauración de espacios de experimentación regulados o esquemas de regulación adaptables, tal como sugieren Desai y Kroll en su estudio "Confía, pero verifica", emergen como opciones viables para garantizar el control y la rendición de cuentas que estas tecnologías necesitan para un despliegue fiable. Esto ha de compaginarse con la urgencia de cultivar un contexto favorable para la innovación ética y de modular la regulación tomando como base datos concretos sobre las consecuencias palpables de estas tecnologías.

La buena gobernanza de los sistemas algorítmicos también requerirá una reconsideración sustantiva de las prácticas organizacionales y profesionales a través de las cuales se desarrollan. La incorporación sistemática de evaluaciones de impacto algorítmico, procesos de diseño participativo y mecanismos sólidos de trazabilidad y documentación es una piedra angular de una estructura institucional que permitiría identificar y mitigar riesgos discriminatorios antes de que resulten en daño social significativo.

Este cambio requiere no solo recursos materiales y conocimientos técnicos específicos, sino fundamentalmente transformaciones culturales profundas de comunidades profesionales clásicamente organizadas en torno a valores como la optimización técnica o la eficiencia computacional. Para poner en práctica los principios abstractos de responsabilidad algorítmica esto también dependerá de capacitar a profesionales técnicos que piensen desde las dimensiones éticas, legales y sociales de su trabajo, y de crear entornos organizacionales que recompensen y valoren estas consideraciones.

Para lograr esto, será imprescindible desarrollar algoritmos que estén en sintonía con los principios esenciales de la democracia, lo que demanda un trabajo conjunto y constante que trascienda las barreras entre disciplinas, industrias y naciones. Es crucial abordarlos desde un enfoque verdaderamente transdisciplinario, donde las visiones técnicas, jurídicas, filosóficas y sociológicas se combinen de manera práctica y coherente, con el fin de crear soluciones que sean a la vez técnicamente factibles, normativamente legítimas e institucionalmente viables.

Este diálogo social profundo y necesario sobre el porvenir de los algoritmos no solo es viable para mitigar los múltiples riesgos asociados con estas innovadoras tecnologías, sino que también resulta esencial para reafirmar los principios democráticos básicos de autonomía colectiva, equidad real y respeto por la dignidad humana como pilares fundamentales del progreso tecnológico actual.

14. Impacto de la IA en el empleo y la economía

Hoy en día, la IA se muestra como una espada de doble filo. Por un lado, se espera que transforme radicalmente las economías, generando nuevos puestos de trabajo y perfeccionando procedimientos que antes requerían mucho tiempo y dinero. Por otro lado, si se aplica sin principios éticos sólidos, podría exacerbar las desigualdades existentes y remodelar o incluso suprimir empleos que sustentan a muchísimas personas. En esta delicada balanza entre avance y vulnerabilidad, es crucial preguntarnos: ¿cómo nos aseguramos de que estas tecnologías no aumenten las diferencias sociales? ¿Cómo impedimos que la eficiencia técnica opaque el valor de la persona?

En su libro, "Automatizar la Desigualdad", Virginia Eubanks nos muestra cómo la IA y el análisis de datos masivos pueden restarles humanidad a las decisiones en el ámbito de los servicios sociales. Un claro ejemplo es el de Sophie Stipes, una joven de Indiana con parálisis cerebral, quien en marzo de 2008 se vio privada de su acceso a *Medicaid* debido a una supuesta falta de colaboración con las autoridades. Si bien recibió una notificación sobre la posible cancelación de su cobertura, Sophie logró recuperar su seguro gracias a la intervención de su madre y de organizaciones comunitarias que llevaron su situación ante el Congreso. Este hecho pone de manifiesto un problema de gran magnitud: un sistema automatizado, implementado por IBM, rechazó más de un millón de solicitudes entre 2006 y 2008, perjudicando a numerosas familias de escasos recursos. Virginia Eubanks, en su obra, subraya cómo la tecnología aplicada en los servicios públicos puede acarrear consecuencias terribles para las personas más vulnerables, dejando al descubierto las decisiones sociales y políticas que subyacen a estos sistemas. Este concepto evidencia la ausencia de sensibilidad en sistemas automatizados que anteponen la eficiencia al apoyo humano, lo que puede impactar negativamente en el bienestar y las oportunidades laborales de quienes ya se encuentran en situaciones precarias. La automatización de decisiones en programas como *Medicaid* no solo perpetúa la desigualdad económica, sino que también fortalece estereotipos dañinos, generando un ciclo de exclusión que afecta negativamente a las comunidades de bajos ingresos. Esto pone de relieve la necesidad de replantearnos cómo se están utilizando las tecnologías en la sociedad, con el fin de garantizar que beneficien a todos y no solo a unos cuantos.

Este problema no solo afecta a las áreas de asistencia social. En su obra "Armas de Destrucción Matemática," Cathy O'Neil (2018) describe cómo firmas como *TechForward* emplean sistemas de contratación basados en algoritmos que, aunque pretenden ser imparciales, acaban por

discriminar a aspirantes capacitados. En este contexto, el sistema automático de selección valoraba a los aspirantes mediante algoritmos que examinaban distintas informaciones. Un postulante, Luis, un programador muy preparado, fue descartado injustamente ya que su domicilio postal fue catalogado como "de alto riesgo. " Dicha clasificación se fundamentó en datos previos que relacionaban zonas económicas precarias con una mayor probabilidad de inconvenientes laborales, como el incumplimiento de horarios o el bajo rendimiento, sin evaluar las aptitudes particulares de los aspirantes.

Cuando los empleados se vieron juzgados por valoraciones basadas en patrones parcializados, la jefa de personal, Clara, comprendió el perjuicio que esos "modelos de aniquilación matemática" podían generar. Cayó en la cuenta de que, si bien esos sistemas auguraban resoluciones imparciales, en verdad estaban afianzando prejuicios y estigmas que perjudicaban de forma desmedida a los empleados de entornos frágiles. Clara comprendió que una visión miope de las estadísticas y los algoritmos no solo dejaba de lado a personas talentosas como Luis, sino que además minaba el ánimo y la unión del equipo en su totalidad. Su firmeza para defender una visión más humana y ecuánime indujo a la empresa a replantearse sus prácticas, enfatizando la necesidad de integrar la ecuanimidad en el manejo de la tecnología. Esta historia revela cómo la automatización, lejos de ser una solución que vale para todo, puede consolidar las desigualdades si no se usa con responsabilidad y moralidad.

Dentro de este contexto, el estudio de Greene, Hoffmann y Stark (2019) enriquece estas explicaciones al examinar de qué manera las proclamaciones de valores relacionadas con el diseño ético de la IA y el Aprendizaje Automático (IA/ML) pueden afectar la forma en que se comprenden y se emplean estas herramientas en el entorno profesional. La investigación sugiere que, a pesar de los esfuerzos por incorporar fundamentos éticos en la creación de la IA/ML, estas herramientas

frecuentemente afianzan y aumentan las disparidades preexistentes. A modo de ejemplo, se enfatiza que muchas de estas afirmaciones asumen que la ética puede tratarse principalmente desde un punto de vista técnico, lo que significa que las resoluciones sobre el diseño y la implementación de los sistemas de IA son tomadas por un grupo limitado de especialistas (Greene et al., 2019). Esto no solo margina a las voces críticas de los sectores perjudicados, sino que también podría resultar en una visión tecnocrática que pasa por alto las consecuencias sociales y económicas más generales.

Por lo tanto, es fundamental que, al analizar el impacto de la IA en el empleo y la economía, se realice un examen crítico que tenga en cuenta no solo los posibles beneficios económicos, sino también las repercusiones sociales y éticas de su uso. La investigación de Greene y sus colegas nos invita a pensar en cómo podemos crear un marco más inclusivo y justo, que priorice el bienestar de todos los trabajadores, en lugar de dejar que las decisiones se basen únicamente en la eficiencia técnica.

Al debatir cómo la IA afecta el trabajo y la economía, es crucial entender que quitar datos delicados no asegura que el sistema sea equitativo. Veale y Binns (2017) señalan que borrar detalles como etnia o sexo puede engañosamente hacer creer que el modelo no tiene sesgos. A pesar de esto, otros factores aparentemente inocuos podrían estar vinculados a esas variables protegidas, manteniendo así la discriminación. Por ejemplo, usar datos geográficos, como los códigos postales, como reemplazo de variables sensibles, puede llevar a deducir características protegidas como la raza, creando así discriminación indirecta.

Para abordar estos desafíos, Veale y Binns proponen tres enfoques que pueden ayudar a mitigar la discriminación en sistemas de aprendizaje automático sin la necesidad de recopilar datos sensibles:

1. Terceros de confianza: Este enfoque sugiere que organizaciones externas, como ONG o entidades gubernamentales, recojan y gestionen datos sensibles. La idea es que estos terceros actúen como intermediarios que pueden recoger información sobre características protegidas, como raza, género o discapacidad, sin que la organización que desarrolla el modelo tenga acceso directo a estos datos. Por ejemplo, un asegurador puede trabajar con una organización de derechos del consumidor para que los clientes proporcionen información sobre características protegidas al momento de adquirir una póliza, permitiendo así evaluar el modelo sin comprometer la privacidad.
2. Plataformas de conocimiento colaborativo: Espacios donde diferentes organizaciones comparten experiencias sobre equidad en sus sistemas de IA, facilitando el aprendizaje de mejores prácticas en la contratación y otros procesos. Por ejemplo, empresas del sector de recursos humanos podrían compartir datos sobre sesgos en sus procesos de selección, ayudando a otras a evitar errores similares.
3. Análisis exploratorio: Uso de técnicas no supervisadas para identificar patrones que puedan indicar sesgo, permitiendo a las organizaciones construir hipótesis sobre características injustas y ajustar sus modelos antes de la implementación. Por ejemplo, un modelo de predicción de riesgo podría utilizar análisis de agrupamiento para identificar subgrupos dentro de sus datos, permitiendo a los analistas observar diferencias de rendimiento y ajustar el modelo en consecuencia.

Estos enfoques no solo buscan mitigar la discriminación, sino que también fomentan una mayor responsabilidad en el uso de la IA en entornos laborales.

En definitiva, el debate sobre IA y empleo no puede reducirse a un cálculo frío de ganancias y pérdidas económicas. Detrás de cada algoritmo hay historias humanas: una madre que pierde su

seguro médico por un error de sistema, un trabajador talentoso etiquetado como "riesgo" por su código postal, o una comunidad entera excluida de oportunidades laborales por prejuicios codificados. La verdadera revolución no estará en la tecnología misma, sino en nuestra capacidad para usarla como herramienta de inclusión, no de marginación. Como sociedad, el desafío es claro: construir un futuro donde la eficiencia no se anteponga a la dignidad, y donde el progreso tecnológico sea sinónimo de equidad.

15. IA, derechos humanos y sostenibilidad

Una de las características que resaltan en el mundo contemporáneo es el reconocimiento de que todo ser humano, por el hecho de serlo, es titular de derechos fundamentales que la sociedad no puede arrebatarle. Estos derechos no dependen de un reconocimiento por el Estado; tampoco dependen de la nacionalidad de la persona ni de la cultura a la cual pertenezca. Son derechos universales que corresponden a todo habitante de la tierra, y se fundamentan en la dignidad inherente a todos los seres humanos. En este sentido, se ha consolidado un marco ético y legal global que sustenta la protección de los derechos humanos, al margen de cualquier sistema político o económico.

Dentro de este panorama, el veloz avance de la IA representa tanto una oportunidad como un riesgo para los derechos fundamentales. Se ha discutido ampliamente cómo la IA está transformando múltiples campos, que incluyen la salud, la educación, la seguridad, la economía y el sistema judicial. No obstante, esta transformación no es imparcial ni ocurre por sí sola, sino que implica decisiones técnicas, políticas y morales que afectan directamente a los individuos. En consecuencia, el desarrollo, la creación y el análisis de los sistemas de IA deben basarse en

principios y defender la igualdad, la justicia social, la sostenibilidad y el respeto por los derechos humanos y la dignidad.

La declaración del 8 de abril de 2019 de la Comisión Europea detalla los Siete requisitos clave para conseguir una IA confiable. Estos siete requisitos son:

1. La intervención y supervisión humanas: los sistemas de IA deben promover una sociedad equitativa, apoyar la intervención humana y los derechos fundamentales, y no disminuir, limitar o desorientar la autonomía humana.
2. Solidez y seguridad técnicas: la IA requiere que los algoritmos sean lo suficientemente seguros, fiables y solidos para resolver errores o incoherencias durante todas las fases del ciclo de vida útil del sistema de IA.
3. Privacidad y gestión de datos: las personas deben tener pleno control sobre sus propios datos, y que los datos que les conciernen no se utilizarán para perjudicarlos o discriminarlos.
4. Transparencia: se debe llevar la trazabilidad de los sistemas de IA.
5. Diversidad, no discriminación y equidad: los sistemas de IA deben tener en cuenta las capacidades, habilidades y necesidades humanas y garantizar la accesibilidad.
6. Bienestar social y medioambiental: la IA debe fomentar la sostenibilidad y la responsabilidad ecológica.
7. Rendición de cuentas: deben implementarse mecanismos que garanticen la responsabilidad y la rendición de cuentas de los sistemas de IA y de sus resultados.

Tales fundamentos actúan como pilares cruciales, asegurando que la IA no infrinja las libertades, sino que, por el contrario, las fortalezca. Tanto la capacidad de rastreo como la gestión por parte de personas son suficientes para una interacción ética con estas herramientas.

En octubre de 2018, la organización *The Public Voice* ratificó los Lineamientos Universales para la IA. De los trece requerimientos, destacaremos el primero, el cual aborda la salvaguarda de los derechos humanos: Emplear sistemas de IA basados en los derechos humanos que capaciten al sistema judicial para respetar, amparar y fomentar los derechos.

A pesar de que todos los derechos humanos pueden verse afectados por los sistemas de IA, es imperativo mantener la trazabilidad de estos. (*The Public Voice*, 2018). Los siguientes cuatro aspectos revisten gran importancia en el marco de los trámites judiciales:

1. Equidad: adoptar sistemas de IA que alcancen objetivos mediante procesos que salvaguarden la equidad y garanticen un acceso inclusivo a la tecnología.
2. No discriminación: evitar aplicaciones sesgadas de sistemas de IA y resultados que reproduzcan, refuercen, perpetúen o agraven la discriminación.
3. Equidad procesal: evaluar las implicaciones de los sistemas de IA para la equidad procesal a lo largo del ciclo de vida del sistema de IA.
4. Protección de datos personales: adoptar sistemas de IA que protejan los datos personales tratados para la administración de justicia.

Los sistemas de IA necesitan diseñarse con un compromiso con la protección completa de los derechos, y especialmente cuando se aplican en entornos judiciales. En tales dominios, la IA no solo influye en los procedimientos de gobernanza sino en la igualdad, libertad e incluso en las vidas de los individuos.

Desde el punto de vista de sostenibilidad, el uso de IA es tanto un desafío como una oportunidad. En el lado positivo, la IA puede ser utilizada para optimizar el uso de energía, mejorar la gestión del transporte público, prevenir la deforestación o anticipar desastres naturales (El País, 2025). En ese sentido, es una herramienta para lograr los Objetivos de Desarrollo Sostenible. Pero

viene con costos ambientales: entrenar modelos de lenguaje grandes, por ejemplo, requiere cantidades masivas de energía.

Especialmente el impacto ecológico de la IA exige una crítica a sus fundamentos digitales. Los centros de datos y la minería de criptomonedas, ambos con un uso elevado de energía y agua, han sido criticados por socavar compromisos ambientales globales. Así, cualquier política pública o empresarial en torno a la IA necesita incluir la huella de carbono y el costo para el ecosistema local de la IA.

Adicionalmente, la UNESCO consolida esta idea en su Recomendación sobre la Ética de la IA (2021), señalando que la puesta en práctica de la IA ha de fundamentarse en valores éticos globales y buscar el beneficio de todos (UNESCO, 2021). Esta directriz sirve de base para que las naciones establezcan normas sobre la IA sin poner en riesgo los derechos fundamentales ni el equilibrio ecológico. Entre los pilares que resalta están la dignidad humana, el cuidado del entorno, la equidad entre generaciones, la integración y la paridad. Asimismo, indica que es imprescindible asegurar la salvaguarda de la información personal, evitar las inclinaciones en los algoritmos y asegurar un manejo responsable y claro de los sistemas de IA.

En Latinoamérica, entidades como Derechos Digitales han puesto de manifiesto los peligros de usar la IA sin la debida vigilancia, sobre todo en lo que respecta a la vigilancia, la seguridad ciudadana o en los algoritmos que influyen en resoluciones judiciales o sociales (Derechos Digitales, s. f.). Esta perspectiva ha aumentado la complejidad que subyace a la batalla por la defensa de la información. En escenarios de gran desigualdad y escasa protección de datos, las tecnologías de reconocimiento facial o los sistemas predictivos de delitos podrían intensificar la marginación social en las naciones.

Derechos Digitales subraya que, ante la ausencia de certezas institucionales y claridad, la implementación de sistemas de IA podría traducirse en el ejercicio abusivo del poder, el quebrantamiento de las libertades ciudadanas y la discriminación hecha por máquinas. Asimismo, plantean la creación de normas legales para directrices regionales que garanticen la vigilancia ciudadana, la rendición de cuentas y el respeto de los derechos fundamentales.

Además de lo anterior, vale la pena considerar cómo la IA podría ser útil para defender los derechos fundamentales de las personas en momentos de crisis. Un caso ilustrativo son las emergencias humanitarias, donde la IA ayuda a detectar áreas de riesgo, a distribuir la asistencia de forma más eficiente o a prever el surgimiento de brotes infecciosos. Estas herramientas son cruciales para las ONGs y las organizaciones globales que trabajan en el campo. No obstante, estos sistemas deben diseñarse protegiendo la privacidad de los individuos más vulnerables, y adhiriéndose a principios éticos rigurosos.

Mientras tanto, la relación entre IA y sostenibilidad se ha fortalecido mediante los sistemas inteligentes dentro en la agricultura de precisión. De esta manera, los agricultores pueden usar menos agua, fertilizantes y pesticidas con un impacto ambiental reducido al usar sensores y algoritmos predictivos. Estas prácticas agrícolas responsables ayudan no solo a cuidar el planeta, sino también a enfatizar el derecho a la alimentación para todos, particularmente en áreas donde el clima está cambiando.

Por otro lado, la IA y la sostenibilidad cada vez están más unidas gracias a la agricultura de precisión. Los agricultores pueden usar menos agua, abonos y pesticidas, lo que reduce el impacto ambiental, utilizando sensores y cálculos predictivos. Estas prácticas agrícolas responsables ayudan a proteger el planeta y aseguran el derecho a la alimentación, sobre todo en zonas afectadas por el cambio climático.

Por otro lado, las ciudades inteligentes (*smart cities*) representan un campo donde la IA y la sostenibilidad convergen. Mediante el uso de algoritmos, sensores y *big data*, las ciudades pueden gestionar de manera más eficiente el tráfico, el alumbrado público, la recolección de basura y el consumo energético. Esto mejora la calidad de vida de los ciudadanos y garantiza el acceso equitativo a servicios públicos, lo cual se relaciona directamente con los derechos humanos.

En el ámbito laboral, la automatización impulsada por IA ha encendido debates sobre el derecho al trabajo digno. Aunque la IA puede aumentar la productividad y generar nuevos empleos, también puede desplazar a los trabajadores, particularmente en sectores expuestos. Por eso los gobiernos progresistas se mueven rápidamente para establecer programas de reentrenamiento laboral y educación continua para que nadie se quede atrás en la transición tecnológica.

16. Conclusiones

Esta investigación nos brindó una comprensión más clara sobre cómo se categoriza y expresa la IA, considerando tanto su aspecto técnico como su influencia en la sociedad. El estudio de sus fundamentos teóricos, sus aplicaciones actuales, los retos que presenta y el estado de su regulación, hizo posible identificar tendencias que reflejan no solo el presente, sino también la posible evolución de esta tecnología. Los resultados confirman que la IA es mucho más que una simple innovación tecnológica; es un fenómeno que se conecta con las estructuras sociales, económicas y políticas ya establecidas, generando cambios profundos en varios aspectos de la vida diaria y las organizaciones.

Durante la investigación, se constató la rápida evolución de la IA, que ha integrado diversas técnicas como el aprendizaje automático y el procesamiento del lenguaje natural. Esto ha posibilitado la creación de sistemas con capacidades que antes solo atribuíamos a la inteligencia humana. En la actualidad, la IA está presente en sectores como la industria, las finanzas, la salud y la educación, transformando no solo los procesos, sino también la interacción de las personas con estas tecnologías. Esta presencia genera un doble impacto: por un lado, mejora la eficiencia y crea nuevas oportunidades; por el otro, plantea desafíos relevantes en cuanto al empleo, la distribución del poder tecnológico y las formas convencionales de interacción social.

El estudio revela de forma inequívoca que las ventajas que brinda la IA son extensas y palpables. En cuanto a la eficiencia, hace posible la optimización de procedimientos intrincados, la disminución de gastos y la mejora de la calidad en diversas industrias. En el ámbito sanitario, por ejemplo, ha favorecido diagnósticos más veloces y certeros; en la enseñanza, ha facilitado la adaptación del aprendizaje a cada estudiante; y en las administraciones públicas, ha ayudado a que

sean más accesibles. Igualmente, al simplificar el acceso a instrumentos de análisis sofisticados, la IA contribuye a la democratización del saber. Además, juega un papel crucial en cuestiones de sostenibilidad, colaborando en la gestión más inteligente de los recursos naturales y en la confrontación de retos medioambientales de forma más innovadora.

A pesar de esto, el progreso de la IA también trae consigo peligros que no se pueden ignorar. Uno de los más evidentes es su repercusión en el mercado laboral: ya no solo se trata de sustituir a los trabajadores, sino de transformar radicalmente el tipo de aptitudes que se necesitan. Otro inconveniente grave son las tendencias discriminatorias en los algoritmos, que pueden intensificar las desigualdades sociales ya existentes. Por otro lado, la acumulación de capacidades avanzadas en unas cuantas empresas genera desequilibrios de poder, perjudicando tanto a la competencia como a la autonomía de los usuarios. La intimidad, la toma de decisiones automática y la dignidad humana también se ven en peligro cuando las personas se reducen a información. Por último, es preciso avanzar en mecanismos de control y transparencia, pues muchos de estos sistemas funcionan como cajas negras complejas de examinar o supervisar con eficacia.

En Colombia, la velocidad con la que avanza la IA contrasta fuertemente con la lentitud de las leyes. Si bien contamos con la Ley 1581 de 2012 para proteger datos, no basta para los desafíos que presenta esta tecnología. La IA opera de manera diferente a los sistemas de siempre, exigiendo normas hechas a su medida que contemplen la dificultad para entender los algoritmos y la independencia de los sistemas. No es un problema solo nuestro; en todo el mundo, las leyes batallan por seguir el paso a la innovación. Además, la regulación debe ser integral, pues la IA afecta áreas tan diversas como la salud, las finanzas, la educación y las comunicaciones. El futuro de esta tecnología dependerá de la participación de todos: gobierno, empresas, universidades y ciudadanos. Es vital que la IA se cree y se aplique pensando en las personas, no para reemplazarlas,

sino para mejorar sus habilidades. Para lograrlo, necesitamos una gestión adaptable, que permita cambios constantes según lo que vayamos aprendiendo. Esto implica crear políticas y principios éticos que no solo acompañen la tecnología, sino que la orienten hacia el bien común.

Es muy probable que la IA se convierta, dentro de los años que vienen, en una herramienta tecnológica que cambie de forma importante nuestra existencia, nuestro empleo y nuestra estructura social. Su impacto será tan fuerte como lo fue en su momento la llegada de la industria o la era digital. Seremos testigos de cómo se integra cada vez más en los sistemas fundamentales de la sociedad: los transportes, la energía, la sanidad, la enseñanza. Esto generará alteraciones importantes en el mundo laboral, en la economía y en la forma en que se entiende el trabajo. Además, será imprescindible tener sistemas de formación que capaciten a los individuos para desempeñar nuevos puestos, con mayor soltura y aptitud para adaptarse.

La IA supone una ocasión inigualable, pero a la vez un reto difícil. Conseguir que sus ventajas alcancen a todos y que sus peligros se manejen de forma sensata dependerá de las elecciones que hagamos en el presente. La respuesta está en impulsar un desarrollo que valore los derechos humanos, que impulse la igualdad y que sea viable a largo plazo. No existe un único sendero ni una meta fija: el porvenir de la IA lo edificaremos entre todos, y el resultado dependerá de si decidimos poner la tecnología al servicio de los ciudadanos, o dejamos que avance sin control. Lo primordial es recordar siempre que la tecnología debe ayudarnos a tener una vida mejor, no sustituir aquello que nos hace personas.

17. Recomendaciones

1. Promover una regulación actualizada y transversal
    - Elaborar reglamentos específicos que tomen en cuenta las características técnicas de los sistemas de IA, tales como su autonomía y capacidad de aprendizaje.
    - Asegurar que la normativa se extienda a áreas como salud, educación, justicia y finanzas, donde la IA ejerce un impacto en aumento.
2. Fortalecer la educación y la formación en IA
    - Incorporar materias relacionadas con la IA, la ética digital y el pensamiento crítico en los planes de estudio desde los niveles iniciales hasta la educación universitaria.
    - Promover la formación técnica y humanística de profesionales para que tengan la capacidad de enfrentar los retos éticos, sociales y laborales que conlleva la IA.
3. Garantizar la inclusión y minimizar brechas
    - Garantizar que las ventajas de la IA sean accesibles para todos los habitantes, en particular para comunidades en situación de vulnerabilidad, evitando su exclusión tecnológica.
    - Elaborar políticas que reduzcan los impactos adversos en el trabajo, fomentando la transformación laboral y nuevas oportunidades de trabajo.
4. Establecer mecanismos de vigilancia y control
    - Establecer entidades o comités autónomos que monitoreen el uso técnico de la IA, detectando peligros y sugiriendo soluciones correctivas.
    - Promover el establecimiento de normas técnicas y buenas prácticas para el desarrollo seguro y fiable de sistemas inteligentes.

Referencias Bibliográficas

ACLU. (2018, julio 26). Amazon's face recognition falsely matched 28 members of Congress with mugshots. (El reconocimiento facial de Amazon coincidió falsamente con 28 miembros del Congreso con fotos policiales). https://www.aclu.org/news/privacy-technology/amazons-face- recognition-falsely-matched-28

Alayón Miranda, S. (2024). El problema de la interpretabilidad de la inteligencia artificial y su impacto en la administración pública. Revista Canaria de Administración Pública, 3, 175-202.

American Psychological Association. (2020). Publication manual of the American Psychological Association (Manual de publicación de la Asociación Americana de Psicología) (7.a ed.).

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety (Problemas concretos en materia de seguridad de la IA). Cornell University, 29. https://doi.org/arXiv:1606.06565v2

Analysis of AI agents: types, capabilities and applications (Análisis de los agentes de IA: tipos, capacidades y aplicaciones.). (2025). Automation Anywhere. https://www.automationanywhere.com/la/company/blog/automation- ai/exploring-ai-agents-types-capabilities-and-real-world-applications.

Angwin, J., Kirchner, L., Larson, J., & Matty, S. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against Blacks (Prejuicios de las máquinas: Hay un software utilizado en todo el país para predecir futuros criminales y está sesgado contra los negros). [Benton Institute for Broadband y

Society].

Annual report on ethical incidents in artificial intelligence systems New York University. (2024). AI Now Institute.

Aprovechar las oportunidades de sistemas seguros, protegidos y fiables de inteligencia artificial para el desarrollo sostenible. (2024). Asamblea General de las Naciones Unidas. https://docs.un.org/es/A/RES/78/265

Argandoña, A. (2019). Ética e inteligencia artificial (I). IESE Economía, Ética y RSE. https://blog.iese.edu/antonioargandona/2019/03/25/etica-e-inteligencia-artificial-i/

Aristoteles. (1992). Capítulo 4 y 5 del libro III (A. Schmidt, Trad.). En De Anima.

Arntz, M., Gregory, T., & Zierahn, U. (2016). The risk of automation for jobs in OECD countries (El riesgo de la automatización para los empleos en los países de la OCDE). OECD Publishing. https://doi.org/10.1787/5jlz9h56dvq7-en

Arrollo, A. (2022). CLIPS, el motor de reglas de la NASA. Adriánistan. https://blog.adrianistan.eu/clips-motor-reglas-nasa/

ATLAS, el controvertido programa con el que Estados Unidos quiere dotar de inteligencia artificial a sus tanques. (2019). BBC News Mundo. https://www.bbc.com/mundo/noticias- 47554998

Bear, M. F., Connors, B. W., & Paradiso, M. A. (2020). Neurociencia: La exploración del cerebro (4.a ed.). Wolters Kluwer.

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim Code (Raza después de la tecnología: Herramientas abolicionistas para el nuevo Código Jim). Polity Press.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles

(El dilema social de los vehículos autónomos). Science, 352(6293), 1573-1576. https://doi.org/10.1126/ciencia.aaf2654

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies (Superinteligencia: Caminos, peligros, estrategias). Oxford University Press.

Brennen, J. S., Simón, F., Howrd, P., & Kleis Nielsen, R. (2020). Types, sources, and claims of COVID-19 misinformation (Tipos, fuentes y afirmaciones de desinformación sobre COVID-19). Oxford University. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation.

Brey, P. (2022). Ethics by design in robotics and AI. Ethics and Information Technology (Ética por diseño en robótica e IA. Ética y Tecnología de la Información), 24(1), 1–15. https://doi.org/10.1007/s10676-021-09568-5

Brey, P., & Dainow, B. (2024). Ethics by design for artificial intelligence (Ética por diseño para la inteligencia artificial). AI and Ethics. https://doi.org/10.1007/s43681-023-00330-4

Broussard, M. (2018). Artificial unintelligence: How computers misunderstand the world (Desinteligencia artificial: Cómo las computadoras malinterpretan el mundo). MIT Press.

Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies (La segunda era de las máquinas: Trabajo, progreso y prosperidad en una época de tecnologías brillantes). W. W. Norton & Company.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research

(Sombras de género: Disparidades de precisión interseccionales en la clasificación comercial de género. Actas de Investigación en Aprendizaje Automático), 81, 1–15. https://proceedings.mlr.press/v81/buolamwini18a.html

California Learning Resource Network. (2025). What is design ethics? (¿Qué es la ética del diseño?) CLRN. https://www.clrn.org/what-is-design-ethics/

Catmull, E. (2014). Creativity, Inc.: Overcoming the unseen forces that stand in the way of true inspiration (Creatividad, S.A.: Superando las fuerzas invisibles que se interponen en el camino de la verdadera inspiración). Random House.

Cavoukian, A. (2011). Privacy by design: The 7 foundational principles. Information and Privacy Commissioner of Ontario (Privacidad desde el diseño: Los 7 principios fundamentales. Comisionado de Información y Privacidad de Ontario).https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security(Deep fakes: Un desafío inminente para la privacidad, la democracia y la seguridad nacional). Harvard Law Review, 132(1), 1–38. https://harvardlawreview.org/

Christian, B. (2020). The alignment problem: Machine learning and human values(El problema de la alineación: Aprendizaje automático y valores humanos). W. W. Norton & Company.

Chun, J., Schroeder de Witt, C., & Elkins, K. (2024). Comparative global AI regulation: Policy perspectives from the EU, China, and the US [Preprint] (Regulación global comparativa de la IA: Perspectivas políticas desde la UE, China y los EE. UU.). arXiv.

https://arxiv.org/abs/2410.21279v1

Coeckelbergh, M. (2020). AI ethics (Ética de la IA). MIT Press.

Comisión Europea. (2019). Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones: Construyendo la confianza en la inteligencia artificial centrada en el ser humano [COM (2019) 168 final]. https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52019DC0168

Cómo la visión por computadora está transformando la industria de la salud. (2022). AISuperior. https://aisuperior.com/es/blog/how-computer-vision-is-transforming-the-healthcare-industry/

Congreso de Colombia. (2008). Ley 1266 de 2008. Por la cual se dictan las disposiciones generales del habeas data y se regula el manejo de la información contenida en bases de datos personales de carácter financiero, crediticio, comercial, de servicios y la proveniente de terceros países. Diario Oficial No. 47.219. https://www.oas.org/es/sla/ddi/docs/CO%2014%20Ley%201266%20Habeas%20Data.pd f

Congreso de Colombia. (2009, 5 de enero). Ley 1273 de 2009. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=34492

Congreso de Colombia. (2012). Ley 1581 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales. Diario Oficial No. 48.587. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

Consejo de Derechos Humanos de las Naciones Unidas. (2021). El derecho a la privacidad en

la era digital (A/HRC/48/31). Naciones Unidas. https://docs.un.org/es/A/HRC/48/31

Consejo de la Unión Europea. (2024). Reglamento de inteligencia artificial. https://www.consilium.europa.eu/es/policies/artificial-intelligence/

Cortina, A. (2024). ¿Ética o ideología de la inteligencia artificial? El eclipse de la razón comunicativa en una sociedad tecnologizada. Paidós.

Costanza-Chock, S. (2020). Design justice: Community-led practices to build the worlds we need(Justicia del diseño: Prácticas lideradas por la comunidad para construir los mundos que necesitamos).MIT Press.

Craigon, P. J., Sacks, J., Brewer, S., Frey, J., Gutierrez, A., Jacobs, N., Kanza, S., Manning, L., Munday, S., Wintour, A., & Pearson, S. (2023). Ethics by design: Responsible research & innovation for AI in the food sector (Ética por diseño: Investigación e innovación responsables para la IA en el sector alimentario). Journal of Responsible Technology, 13, 100051. https://doi.org/10.1016/j.jrt.2022.100051

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence(Atlas de la IA: Poder, política y los costos planetarios de la inteligencia artificial). Yale University Press.

De Vries, G. (2024, septiembre 12). EUROfusion lidera los avances en inteligencia artificial y aprendizaje automático para liberar la energía de fusión. EUROfusion. https://euro- fusion.org/eurofusion-news/eurofusion-spearheads-advances-in-artificial-intelligence- and-machine-learning-to-unlock-fusion-energy/

Derechos Digitales. (s.f.). Inteligencia artificial. https://www.derechosdigitales.org/tematica/inteligencia-artificial/

Desai, D. R., & Kroll, J. A. (2017). Trust but verify: A guide to algorithms and the law (Confía pero verifica: Una guía sobre algoritmos y la ley). Harvard Journal of Law & Technology.

Díaz-Ramírez, J. (2021). Aprendizaje automático y aprendizaje profundo. Ingeniare: Revista Chilena de Ingeniería, 29(2), 182–183.

Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way(Inteligencia artificial responsable: Cómo desarrollar y usar la IA de manera responsable). Springer. https://doi.org/10.1007/978-3-030-30371-6

Dignum, V., et al. (2018). Ethics by design: Necessity or curse? (Ética por diseño: ¿Necesidad o maldición?) En Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 60–66).

El País. (2020, febrero 12). Un tribunal holandés prohíbe el uso de un sistema de inteligencia artificial para detectar fraudes sociales. El País. https://elpais.com/tecnologia/2020/02/12/actualidad/1581512850_757564.html

El País. (2025, enero 26). Cuando la IA se pone del lado del desarrollo sostenible: La tecnología no nos salvará, pero puede ayudar. El País. https://elpais.com/america-futura/2025-01- 26/cuando-la-ia-se-pone-del-lado-del-desarrollo-sostenible-la-tecnologia-no-nos-salvara- pero-puede-ayudar.html

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor(Automatizando la desigualdad: Cómo las herramientas de alta tecnología perfilan, vigilan y castigan a los pobres). St. Martin's Press.

Comisión Europea. (2020). SIENNA: Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact(SIENNA: Ética informada por las

partes interesadas para nuevas tecnologías con alto impacto socioeconómico y en los derechos humanos). https://sienna-project.eu/

Comisión Europea. (2021). Ethics by design and ethics of use approaches for artificial intelligence(Enfoques de ética por diseño y ética de uso para la inteligencia artificial) (Horizon Europe). https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021- 2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial- intelligence_he_en.pdf

Eyal, N. (2017, junio 26). Here's how Amazon's Alexa hooks you. A four-step model explains the psychology behind what makes technology so habit-forming(Así es como Amazon Alexa te atrapa. Un modelo de cuatro pasos explica la psicología detrás de lo que hace que la tecnología sea tan adictiva). Inc. https://www.inc.com/nir-eyal/heres-how-amazonsalexa-hooks-you.html

Floridi, L. (2018). Soft ethics and the governance of the digital(Ética blanda y la gobernanza de lo digital). Philosophy & Technology, 31(1), 1–8.

Floridi, L. (2019). The ethics of artificial intelligence: Issues and initiatives(La ética de la inteligencia artificial: Problemas e iniciativas). European Parliament Research Service. https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)634452

Ford, M. (2015). Rise of the robots: Technology and the threat of a jobless future(El auge de los robots: La tecnología y la amenaza de un futuro sin Empleo). Basic Books.

Foster, D. (2023). Generative deep learning. O'Reilly Media.

Future of Life Institute. (2017). Asilomar AI principles(Principios de Asilomar sobre IA). https://futureoflife.org/open-letter/ai- principles/

Gamboa, S. C. (2024). Epistemología de la tecnología. Universidad Industrial de Santander.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., &

Crawford, K. (2018). Datasheets for datasets(Hojas de datos para conjuntos de datos). arXiv preprint arXiv:1803.09010. https://arxiv.org/abs/1803.09010

Geostrategy. (2022, octubre 12). Visión artificial en la medicina. https://www.geostrategydata.com/vision-artificial-en-la-medicina/

VASS Latam. (2024, septiembre 10). Gobierno inteligente: Inteligencia artificial en la administración pública. https://vasscompany.com/latam/es/insights/blogs-articles/gobierno-inteligente/

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning(Mejor, más agradable, más claro, más justo: Una evaluación crítica del movimiento por una inteligencia artificial y un aprendizaje automático éticos).

Hicks, M. (2019, septiembre 3). Fraudsters deepfake CEO's voice to trick manager into transferring $243,000 (Los estafadores usan un deepfake de la voz del CEO para engañar al gerente y que transfiera dinero.). The Next Web. https://thenextweb.com/news/fraudsters-deepfake-ceos-voice- to-trick-manager-into-transferring-243000

IEEE. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems(Diseño éticamente alineado: Una visión para priorizar el bienestar humano con sistemas autónomos e inteligentes). https://standards.ieee.org/content/ieee- standards/en/industry-

connections/ec/autonomous-systems.html

Jacobs, N. (2021, octubre 6). How 'ethics by design' could remedy the concerns arising from behavior change technologies(Cómo la 'ética por diseño' podría remediar las preocupaciones surgidas por las tecnologías de cambio de comportamiento). Eindhoven University of Technology. https://www.tue.nl/en/news-and-events/news-overview/06-10-2021-how-ethics-by-design-could-remedy-the-concerns-arising-from-behavior-change-technologies

Jasanoff, S. (2016). The ethics of invention: Technology and the human future(La ética de la invención: Tecnología y el futuro humano). W. W. Norton & Company.

Jurafsky, D., & Martin, J. H. (2020). Speech and language processing (Procesamiento del habla y el lenguaje) (3.ª ed.). Pearson.

Kant, I. (2005). Crítica de la razón pura (M. García Morente, trad.). Editorial Tecnos. (Original publicado en 1781/1787)

Korlakunta, S. (2023, enero 29). Deep learning cat and dog classification using TensorFlow(Clasificación de gatos y perros con aprendizaje profundo usando TensorFlow) . Medium. https://korlakuntasaikamal10.medium.com/deep-learning-cat-and-dog- classification-using-tensorflow-8011596d8f96

La sorprendente y poco conocida historia de Eliza, el primer bot conversacional de la historia. (2018). BBC News Mundo. https://www.bbc.com/mundo/noticias-44290222

LeCun, Y., Bengio, Y., & Haffner, P. (2015). Gradient-based learning applied to document recognition (Aprendizaje basado en gradientes aplicado al reconocimiento de documentos) . Proceedings of the IEEE, 86(11), 2278–2324. https://doi.org/10.1109/5.726791

Maldonado, C. E. (2023). Inteligencia artificial digital. Editorial Universidad del Rosario.

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing (Fundamentos del procesamiento estadístico del lenguaje natural). MITPress.

Marín García, S. (2019). Ética e inteligencia artificial. Cuadernos de la Cátedra Caixabank de Responsabilidad Social Corporativa.

McLennan, S., Fiske, A., Celi, L. A., Müller, R., Harder, J., Ritt, K., Haddadin, S., & Buyx, A. (2020). An embedded ethics approach for AI development (Un enfoque de ética integrada para el desarrollo de IA). Nature Machine Intelligence, 2(9), 488-490.

Ministerio de Ciencia, Tecnología e Innovación. (2021). Marco ético para la inteligencia artificial en Colombia. https://minciencias.gov.co/sites/default/files/marco-etico-ia-colombia- 2021.pdf

Mullainathan, S., & Kleinberg, J. (2018). Discrimination in the age of algorithms. Journal of Legal Analysis.

Müller, V. C. (2020). Ethics of artificial intelligence (Ética de la inteligencia artificial). En E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). Metaphysics Research Lab, Stanford University. https://philpapers.org/archive/MLLEOA-5.pdf

Newell, A., & Simon, H. (1972). Human problem solving. Prentice Hall.

Netherlands AI Coalition. (2020). Responsible data sharing in AI [White paper]. https://nlaic.com/wp-content/uploads/2020/10/Responsible-data-sharing-in-AI.pdf

Nelson, R. K., Winling, L, et al. (2023). Mapping Inequality: Redlining in New Deal America (Mapeando la Desigualdad: La Línea Roja en la América del New Deal). Digital Scholarship Lab. https://dsl.richmond.edu/panorama/redlining.

National Institute of Standards and Technology (NIST). (2023). AI risk management framework. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism (Marco de gestión de riesgos de IA. Noble, S. U. (2018). Algoritmos de opresión: Cómo los motores de búsqueda refuerzan el racismo) . NYU Press.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy (Armas de destrucción matemática: Cómo los grandes datos aumentan la desigualdad y amenazan la democracia). Crown Publishing Group.

Organisation for Economic Co-operation and Development (OECD). (2019). Artificial intelligence in society (La inteligencia artificial en la Sociedad). OECD Publishing. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60423

Organisation for Economic Co-operation and Development (OECD). (2019). Recommendation of the Council on Artificial Intelligence (Recomendación del Consejo sobre Inteligencia Artificial (OECD/LEGAL/0449). OECD Publishing. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Ortiz, E., Loyola, R. M., van der Mersch, B. R., Huerta Romo, J. A., & Morfín Rojas, J. A. (2023). Ética en la inteligencia artificial. En El poder de internet y la industria digital: Usos, abusos y buenas prácticas (pp. 18–23). Universidad Iberoamericana.

Pajares Martinsanz, G., & Cruz García, J. M. de la. (2008). Visión por computador: Imágenes digitales y aplicaciones (2ª ed.). Alfaomega.

Parlamento Europeo y Consejo de la Unión Europea. (2024). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo de 13 de marzo de 2024 por el que se establecen normas armonizadas en materia de inteligencia artificial y se modifican determinados

actos legislativos de la Unión (Reglamento de Inteligencia Artificial). Diario Oficial de la Unión

Europea, L 168, 1–157. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024R1689

Parra, S. (2024, 10 de octubre). La revolución de la IA en la medicina: AlphaFold y el Nobel de Química 2024.National Geographic. https://www.nationalgeographic.com.es/ciencia/revolucion-ia-medicina-alphafold-y-nobel-quimica-2024_23415

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information(La sociedad de la caja negra: Los algoritmos secretos que controlan el dinero y la información). Harvard University Press.

Platón. (1996). Timeo (C. García Gual, Trad.). Alianza Editorial. (Texto original c. 360 a.C.)

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaMantia, A.-S., Mooney, R. D., &

White, L. E. (2018). Neuroscience (6ª ed.) (Neurociencia). Oxford University Press.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners(Los modelos de lenguaje son aprendices multitarea no supervisados.). OpenAI.

Rawat, S., Meena, S., & Gupta, P. (2018). Ethical principles in artificial intelligence systems (Principios éticos en los sistemas de inteligencia artificial).Journal of Technology Ethics, 12(2), 105-118.

Red Seguridad. (2025, 20 de enero). Estafa deepfake multimillonaria: El mayor timo digital de la historia. Red Seguridad.

https://www.redeseguridad.com/noticias/ciberseguridad/estafa-deepfake-multimillonaria-el-mayor-timo-digital-de-la-historia/

Reino Unido. (2006). Fraud Act 2006 (Ley de Fraude de 2006). https://www.legislation.gov.uk/ukpga/2006/35

Renza, D., & Ballesteros, D. M. (2023). Fundamentos de visión por computador utilizando aprendizaje profundo. Editorial Redipe.

Repositorio Universidad El Bosque. (n.d.). La visión computacional, una interfaz entre la seguridad del paciente y la tecnología. https://repositorio.unbosque.edu.co/bitstreams/2544f66f-b17b-4792-9e32-0d9ceeb293f9/download

Cumbre sobre la Ética de la Inteligencia Artificial en América Latina y el Caribe. (2024). Roadmap for ethical artificial intelligence for Latin America and the Caribbean 2024–2025.

Roose, K. (2024, 24 de octubre). ¿Se puede culpar a la IA del suicidio de un adolescente? The New York Times en Español. https://www.nytimes.com/es/2024/10/24/espanol/ciencia-y- tecnologia/ai-chatbot-suicidio.html

Russell, S. (2019). Human compatibility: Artificial intelligence and the problem of control (Compatibilidad humana: La inteligencia artificial y el problema del control). Viking.

Russell, S., & Norvig, P. (2010). Artificial intelligence: A modern approach (Inteligencia artificial: Un enfoque moderno). Pearson Education.

Russell, S., & Norvig, P. (2022). Artificial intelligence: A modern approach (Inteligencia artificial: Un enfoque moderno). Pearson.

Sandoya Yépez, L. K., & Mawyin Peña, M. D. (2022). La inteligencia artificial y su impacto en la gestión pública. LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades, 6(1), 710–718. https://doi.org/10.56712/latam.v6i1.3373

Schönherr, F. (2023). Identifying and overcoming organization and ethical barriers to AI adoption (Identificación y superación de barreras organizativas y éticas para la adopción de la IA). University of St. Gallen. https://www.alexandria.unisg.ch/entities/publication/27ff1040- 729a-495b-9831-0d5f4636fa8b

Sherpa.ai. (2023). Sherpa.ai anuncia una plataforma de inteligencia artificial aplicada a la privacidad de datos. SPRI. https://www.spri.eus/es/teics-comunicacion/sherpa-ai-anuncia- una-plataforma-de-inteligencia-artificial-aplicada-a-la-privacidad-de-datos/

Smith, A. M. (2001). Alhacen's theory of visual perception: A critical edition, with English translation and commentary, of the first three books of Alhacen's De aspectibus, the medieval Latin version of Ibn al-Haytham's Kitāb al-Manāẓir (La teoría de la percepción visual de Alhacen: Una edición crítica, con traducción al inglés y comentario, de los primeros tres libros del De aspectibus de Alhacen, la versión latina medieval del Kitāb al-Manāẓir de Ibn al-Haytham.). American Philosophical Society.

Solanki, A., García-Martínez, R., & Gupta, A. (2022). Mapping human values to ethical AI principles: A systematic approach (Mapear los valores humanos a los principios éticos de la IA: Un enfoque sistemático). AI Ethics Journal, 3(1), 45-67.

Stevens, L. (1984). Artificial intelligence: The search for the perfect machine(Inteligencia artificial: La búsqueda de la máquina perfecta). Hayden Book Company. https://archive.org/details/artificialintell0000stev/page/40/mode/2up

Sucar, L. E., & Gómez, G. (2000). Visión computacional. Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE).

Szeliski, R. (2011). Computer vision: Algorithms and applications(Visión por computadora: Algoritmos y aplicaciones). Springer.

The Public Voice. (2018). Directrices universales para la inteligencia artificial. https://thepublicvoice.org/AI-universal-guidelines/

The White House. (2022, octubre). Blueprint for an AI Bill of Rights: Making automated systems work for the American people (Plan para una Declaración de Derechos de la IA: Hacer que los sistemas automatizados trabajen para el pueblo estadounidense). Executive Office of the President of the United States. https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

TechEthos – Future, technology, ethics. All European Academies (Todas las Academias Europeas). (2023). Allea. https://allea.org/techethos-future-technology-ethics/

Trustworthy artificial intelligence in the Asia-Pacific region. (2023). AI Asia Pacific Institute

UNESCO. (2021, 25 de noviembre). Recomendación sobre la ética de la inteligencia artificial. https://www.unesco.org/es/articles/recomendacion-sobre-la-etica-de-la-inteligencia- artificial

UNESCO. (2022). Recomendación sobre la ética de la inteligencia artificial. Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa

UNESCO. (2024, 14 de octubre). Autoridades de 24 países participarán en el 1.er Foro de Altas Autoridades sobre la Ética de la

Inteligencia Artificial. https://www.unesco.org/es/articles/autoridades-de-24-paises-participaran-en-el-1er-foro- de-altas-autoridades-sobre-la-etica-de-la

Unión Europea. (2024). AI Act: Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Ley de IA: Reglamento del Parlamento Europeo y del Consejo que establece normas armonizadas sobre inteligencia artificial). European Commission.

Universidad de Navarra. (n.d.). El reto de la inteligencia artificial para la seguridad y defensa. https://www.unav.edu/web/global-affairs/el-reto-de-la-inteligencia-artificial-para-la- seguridad-y-defensa

U.S. Equal Employment Opportunity Commission. (2023). Guidance on algorithmic fairness in employment(Orientación sobre la equidad algorítmica en el Empleo). https://www.eeoc.gov

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems (La atención es todo lo que necesitas. Avances en Sistemas de Procesamiento de Información Neural), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society(Aprendizaje automático más justo en el mundo real: Mitigando la discriminación sin recopilar datos sensibles. Big Data y Sociedad). https://doi.org/10.1177/2053951717743530

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs algorithmic support in high-stakes public sector decision-making (El diseño de equidad y responsabilidad necesita apoyo algorítmico en la toma de decisiones del sector público en situaciones de alto riesgo). Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.

Vélez Serrano, J. (2003). Visión por computador. Dykinson. https://elibro-net.bibliotecavirtual.uis.edu.co/es/lc/uis/titulos/104894

Yeung, K., & Lodge, M. (2019). Algorithmic regulation(Regulación algorítmica). Oxford University Press.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power(La era del capitalismo de vigilancia: La lucha por un futuro humano en la nueva frontera del poder). PublicAffairs.

Apéndices

Apéndice A. *Protocolo relatoría sesión 1.*

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
INGENIERÍA DE SISTEMAS E INFORMÁTICA

SEMINARIO DE INVESTIGACIÓN
ÉTICA E INTELIGENCIA ARTIFICIAL

Sesión: 1
Tema: Capítulo 3. Fundamentos y vertientes de la Inteligencia Artificial
Fecha: 12 de febrero de 2025

1. Asistencia: funciones y responsables de la sesión

   Director: Sonia Cristina Gamboa
   Relator: Yonathan Camilo Benítez Mancipe
   Protocolante: Andrés Felipe Cárdenas Zárate
   Participantes:
   - Juan Pablo Arias Remolina
   - Sonia Marcela Granados Moreno
   - Neyder Fabian Mosquera Niño

2. **Objetivos de la relatoría:** Establecer las bases para el Capítulo 3 examinando críticamente avances iniciales para fortalecer la claridad conceptual, la estructura argumentativa y el rigor científico del trabajo. Además, se buscó orientar el desarrollo del contenido asegurando una escritura precisa, el uso apropiado de referencias conforme a las normas APA, fundamentos teóricos sólidos y la inclusión de ejemplos actualizados que conectan los fundamentos de la IA con sus aplicaciones prácticas.

3. **Fuentes de información**:

Foster, D. (2023). Generative Deep Learning. Sebastopol, California, EE. UU.: O'Reilly Media, Inc.

Newell, A., & Simon, H. (1972). Human Problem Solving. Englewood Cliffs, Nueva Jersey, EE. UU.: Prentice Hall.

Parra, S. (2024, octubre 10). La revolución de la IA en la medicina: AlphaFold y el Nobel de Química 2024. National Geographic, págs. https://www.nationalgeographic.com.es/ciencia/revolucion-ia-medicina-alphafold-y-nobel-quimica-2024_23415.

Russell , S., & Norvig, P. (2022). Artificial Intelligence: A Modern Approach. Harlow, United Kingdom: Pearson.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control . Nueva York, EE. UU.: Viking .

Russell, S., & Norvig, P. (2010). Artificial Intelligence: A Modern Approach. Upper Saddle River, Nueva Jersey, EE. UU.: Pearson Education, Inc.

Korlakunta, S. (2023, enero 29). Deep learning cat and dog classification using TensorFlow. Medium. https://korlakuntasaikamal10.medium.com/deep-learning-cat-and-dog-classification-using-tensorflow-8011596d8f96

BBC News Mundo. (2018, 3 de junio). La sorprendente y poco conocida historia de Eliza, el primer bot conversacional de la historia. BBC. https://www.bbc.com/mundo/noticias-44290222

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Arrollo, A. (2022, octubre 27). CLIPS, el motor de reglas de la NASA. Adriánistan. https://blog.adrianistan.eu/clips-motor-reglas-nasa/

Stevens, L. (1984). Artificial intelligence: The search for the perfect machine. Hayden Book Company. https://archive.org/details/artificialintell0000stev/page/40/mode/2up

De Vries, G. (2024, 12 de septiembre). EUROfusion lidera los avances en Inteligencia Artificial y Aprendizaje Automático para liberar la energía de fusión. EUROfusion. https://euro-fusion.org/eurofusion-news/eurofusion-spearheads-advances-in-artificial-intelligence-and-machine-learning-to-unlock-fusion-energy/

4. Desarrollo del tema (Síntesis de la relatoría)

4.1 Introducción al capítulo o contenido trabajado: El capítulo 3 presenta los fundamentos conceptuales, filosóficos y técnicos que sustentan el surgimiento y desarrollo de la IA. Comienza con una reseña histórica y considera las contribuciones de disciplinas científicas como la filosofía, la lógica, la cibernética y la informática para comprender cómo surgieron las ideas que sustentan las tecnologías inteligentes actuales. La relatoría planteó un enfoque estructurado de los aspectos simbólicos, conexionistas y generativos de la IA y analiza sus principios, aplicaciones y limitaciones.

4.2 Temas centrales abordados:

- Referencias históricas de la IA: pensamiento lógico, dualismo y materialismo.

- Orígenes de la era moderna: cibernética, expertos en sistemas y la prueba de Turing.
- Aspectos de la IA:
  - IA simbólica: basada en lógica formal y manipulación de símbolos.
  - IA conexionista: inspirada en las redes neuronales del cerebro humano.
  - IA generativa: modelos que pueden crear contenido nuevo basándose en una parte de los datos entrenados.
- Reflexión crítica sobre los retos de la interpretación, la ética y el control en sistemas inteligentes.
- Los argumentos y términos son claros y no se proporcionan ejemplos.

4.3 Conclusiones Preliminares: El capítulo establece las bases conceptuales e históricas de la IA a través de tres vertientes principales: simbólica, conexionista y generativa. Se evidencia que la IA ha sido moldeada por corrientes filosóficas y científicas desde sus orígenes y que, aunque ha alcanzado altos niveles de desempeño técnico, aún no posee comprensión ni conciencia humana. La docente recomendó mejorar la estructura del texto, profundizar los contenidos, citar correctamente con base en fuentes académicas, y acompañar los conceptos con ejemplos claros y actuales.

5. Desarrollo de la discusión grupal

5.1 Contribuciones relevantes de los participantes

- Juan Pablo destacó la importancia de combinar tendencias filosóficas con avances tecnológicos reales para contextualizar mejor los orígenes de la IA.
- Marcela propuso mejorar la estructura del capítulo dividiéndolo claramente en vertientes (simbólica, conexionista y generativa) y acompañar cada uno de los casos reales.
- Neyder sugirió acompañar conceptos teóricos de ejemplos prácticos y aplicados en áreas como salud, industria o educación.

5.2 Preguntas formuladas:

- ¿Con cuáles se dio inicio y qué alcance tenían esos algoritmos?
- ¿Hoy cuáles algoritmos serían ejemplo de estas vertientes?

5.3 Diferentes contrapuntos o interpretaciones: Se debatió sobre la profundidad teórica del capítulo: algunos plantearon que debía ser más preciso y directo, mientras que otros defendieron un desarrollo más amplio para sustentar mejor las aplicaciones prácticas. También hubo diferencias respecto al tipo de fuentes, contrastando entre

la preferencia por trabajos científicos rigurosos y artículos académicos más accesibles.

5.4 Comentarios de la docente: La docente enfatizó la necesidad de una redacción clara, precisa y sin redundancias. Destacó que las referencias deben seguir estrictamente el formato APA, dando prioridad a los libros y artículos científicos. Destacó que los ejemplos y aplicaciones son esenciales para evitar textos puramente teóricos. Deben ser reales, relevantes y estar conectados a la realidad del lector.

Anexos
Video de la sesión: 12-FEB-2025.mp4

Apéndice B. *Protocolo relatoría sesión 2*

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
INGENIERÍA DE SISTEMAS E INFORMÁTICA

SEMINARIO DE INVESTIGACIÓN
ÉTICA E INTELIGENCIA ARTIFICIAL

Sesión: 2
Tema: Capítulo 6. Fundamentos de ética aplicada a la IA y Capítulo 7. Ética y aplicación en la IA.
Fecha: 19 de febrero de 2025

1. Asistencia: funciones y responsables de la sesión

   Director: Sonia Cristina Gamboa
   Relatores: Juan Pablo Arias Remolina y Sonia Marcela Granados Moreno
   Protocolante: Yonathan Camilo Benítez Mancipe
   Participantes:
   - Andrés Felipe Cárdenas Zárate
   - Neyder Fabian Mosquera Niño

2. Objetivos de la relatoría:

   Analizar los capítulos 6 y 7 del seminario, que abordan:
   - Capítulo 6: Fundamentos de ética aplicada a la IA
   - Capítulo 7: Ética y aplicación en la IA

3. Fuentes de información:

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor
Future of Life Institute (2017). Asilomar AI Principles
Gamboa, S. C. (2024). Epistemología de la tecnología
Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo
Maldonado, C. E. (2023). Inteligencia artificial digital
Marín García, S. (2019). Ética e inteligencia artificial
Cortina, A. (2024). ¿Ética o ideología de la inteligencia artificial?

4. Desarrollo del tema (Síntesis de la relatoría)

4.1 Introducción al capítulo o contenido trabajado: El propósito de la relatoría fue analizar los capítulos 6 y 7 del seminario, que tratan sobre los fundamentos éticos de la IA y su traducción a contextos prácticos. Basándonos en diversas fuentes teóricas y normativas, hemos intentado identificar los principios fundamentales del desarrollo e implementación de tecnologías inteligentes, así como los desafíos y tensiones éticas en este ámbito. Nos basamos en referencias como "Automating Inequality" de Virginia Eubanks, "Asilomar AI Principles" del Future of Life Institute, textos académicos como los de Sonia Gamboa, Carlos Eduardo Maldonado y Adela Cortina, y el Reglamento (UE) 2024/1689 del Parlamento Europeo.

4.2 Temas centrales abordados:

Fundamentos éticos de la IA:

- La relación entre ética e IA
- La no neutralidad de la IA
- La responsabilidad de los desarrolladores
- Los retos éticos en aplicaciones prácticas

Principios éticos fundamentales:

- Respeto a la autonomía humana
- Transparencia
- Responsabilidad y rendición de cuentas
- Robustez y seguridad
- Justicia y no discriminación

Métodos de implementación:

- Técnicos (Ethics by design, IA explicable, Prueba y validación)
- No técnicos (Regulación, Certificaciones, Educación, Investigación)

4.3 Conclusiones Preliminares: La ética en la IA es esencialmente la ética de quienes la desarrollan. Por tanto, es esencial encontrar un equilibrio entre la innovación tecnológica y el respeto de los principios éticos fundamentales. Para lograr este objetivo se necesita un marco regulatorio claro y actualizado para su desarrollo y aplicación. La transparencia y la rendición de cuentas también deben ser pilares fundamentales a lo largo de todo el ciclo de vida de los sistemas de IA. Por último, es esencial prevenir cualquier forma de discriminación y garantizar la protección de los grupos más vulnerables frente a posibles efectos negativos.

5. Desarrollo de la discusión grupal

5.1 Contribuciones relevantes de los participantes: Durante la discusión grupal, se hicieron contribuciones significativas. Se enfatizó la necesidad de respaldar las opiniones con literatura relevante y se resaltaron los términos que requieren una mayor definición. También se discutieron los principios de la ética kantiana y su aplicabilidad a los sistemas autónomos.

5.2 Preguntas formuladas: Las preguntas se centraron en los límites del alcance de la IA, la distribución de responsabilidades en su desarrollo, los principales desafíos éticos en este campo y los parámetros que deben guiar las decisiones respecto a los sistemas autónomos. Los debates también se centraron en la autonomía de la IA respecto de la responsabilidad humana, la dificultad de garantizar la transparencia en sistemas complejos, la tensión entre eficiencia y derechos, y la evaluación de riesgos y beneficios en diferentes contextos de aplicación.

5.3 Diferentes contrapuntos o interpretaciones: En el debate se plantearon contrastes clave entre la autonomía de los sistemas de IA y la responsabilidad humana. Se enfatizó que los humanos siempre deben ser responsables de las decisiones automatizadas. También se ha planteado la dificultad de garantizar la transparencia en sistemas complejos como las redes neuronales, ya que esto compromete la confianza y la explicabilidad. Otro punto importante fue el conflicto entre la eficiencia tecnológica y la protección de los derechos. Se destacó que la búsqueda de la optimización no debe comprometer ni la justicia ni la privacidad.

5.4 Comentarios de la docente: La docente hizo recomendaciones importantes respecto a la presentación del trabajo, incluyendo el uso de un formato conciso, un estilo formal y académico y una clara distinción entre citas directas y argumentos. También sugirió resaltar términos en otros idiomas en cursiva y la posibilidad de fusionar los capítulos 6 y 7 debido a su similitud temática. Por último, destacó que el texto debe ser comprensible incluso para los no expertos en la materia, manteniendo la precisión científica.

Anexos

Video de la sesión: 19-FEB-2025.mp4

Apéndice C. *Protocolo relatoría sesión 3*

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
INGENIERÍA DE SISTEMAS E INFORMÁTICA

SEMINARIO DE INVESTIGACIÓN
ÉTICA E INTELIGENCIA ARTIFICIAL

Sesión: 3
Tema: Capítulo 8. regulación y políticas mundiales sobre IA
Fecha: 26 de febrero de 2025

1. Asistencia: funciones y responsables de la sesión

   Director: Sonia Cristina Gamboa
   Relator: Yonathan Camilo Benítez Mancipe
   Protocolante: Juan Pablo Arias Remolina
   Participantes:
   - Andrés Felipe Cárdenas Zárate
   - Sonia Marcela Granados Moreno
   - Neyder Fabian Mosquera Niño

2. **Objetivos de la relatoría**: Examinar la aplicación de la IA en la vida cotidiana y su regulación. El foco se centró en el impacto de esta tecnología en diversas áreas de la sociedad, como la salud, la economía y la política, así como en las preocupaciones éticas y regulatorias asociadas**.**

3. **Fuentes de información**: Las fuentes utilizadas incluyeron regulaciones internacionales y leyes de Estados Unidos sobre IA, así como referencias culturales como la serie Black Mirror de Netflix y casos de estudio como el sistema de crédito social implementado en China, permitiendo así un análisis multidisciplinar que combina enfoques legales, tecnológicos y socioculturales.

4. Desarrollo del tema (Síntesis de la relatoría)

4.1 Introducción al capítulo o contenido trabajado: La sesión abordó la aplicación de la IA en diversos ámbitos de la sociedad, el papel de los tecnócratas en el poder (como Elon Musk y Donald Trump), las nuevas formas de formulación de políticas a través de la IA y la posibilidad de desarrollar niveles de conciencia a través de la IA**.**

4.2 Temas centrales abordados:

Conciencia y evolución en IA:
- Debate sobre la superación de la programación original
- Posibilidad de conciencia en las máquinas
- Cumplimiento normativo

Regulaciones y Políticas Mundiales sobre IA:

Clasificación de riesgos inaceptables:
- Manipulación psicológica
- Reconocimiento facial no autorizado
- Sistemas de puntaje social
- Reconocimiento de emociones

Regulaciones por países:

Estados Unidos:
- Supervisión humana obligatoria
- Protección de datos
- Directrices sectoriales

4.3 Conclusiones Preliminares:
- La IA tiene un impacto profundo en la vida cotidiana, política y economía global
- Es necesario establecer regulaciones claras
- Se requieren auditorías y supervisión efectiva
- La discusión sobre el papel de la IA debe ser continua

5. Desarrollo de la discusión grupal

5.1 Contribuciones relevantes de los participantes: Durante la discusión grupal, los participantes brindaron información valiosa sobre la diferencia entre la evolución y la adaptación de las máquinas, analizaron las diferencias regulatorias entre países y discutieron el papel de la extrema derecha en el control y gobierno de la IA. También se abordó el acceso desigual a esta tecnología y el control sobre ella.

5.2 Preguntas formuladas:

- ¿Las máquinas están realmente superando su programación original?
- ¿Puede considerarse esto una evolución o simple adaptación?
- ¿Cómo asegurar que la IA respete las regulaciones impuestas?
- ¿Cómo evitar consecuencias inesperadas?

5.3 Diferentes contrapuntos o interpretaciones: Los contrapuntos incluyeron debates sobre la autonomía de la IA versus el control humano, el equilibrio entre beneficios, riesgos y las diferencias regulatorias globales. La estructura de la sesión permitió una discusión amplia y crítica, con tiempo para preguntas y reflexiones finales.

5.4 Comentarios de la docente: La docente incentivó a la comparación entre diferentes marcos regulatorios, así como la reflexión sobre los riesgos éticos y sociales de la IA, especialmente en contextos políticos sensibles. También resaltó la importancia de integrar referencias culturales, como la serie Black Mirror, para enriquecer el análisis desde una perspectiva más cercana al día a día de los estudiantes. En cuanto a la forma, la docente exigió el cumplimiento de ciertos criterios académicos, como justificar los argumentos con fuentes bibliográficas, utilizar lenguaje técnico y promover el respeto por distintas posturas dentro del debate.

Anexos

Video de la sesión: 5-MAR-2025.mp4

Apéndice D. *Protocolo relatoría sesión 4*

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
INGENIERÍA DE SISTEMAS E INFORMÁTICA

SEMINARIO DE INVESTIGACIÓN
ÉTICA E INTELIGENCIA ARTIFICIAL

Sesión: 4
Tema: Capítulo 9. Ética en el diseño y desarrollo de sistemas de IA, Capítulo 10. IA en la toma de decisiones y gobernanza algorítmica y Capítulo 11. Impacto de la IA en el empleo y la economía
Fecha: 05 de marzo de 2025

1. **Asistencia: funciones y responsables de la sesión**
   Director: Sonia Cristina Gamboa
   Relatores: Juan Pablo Arias Remolina, Neyder Fabian Mosquera Niño y Andrés Felipe Cárdenas Zárate
   Protocolante: Sonia Marcela Granados Moreno
   Participantes: Yonathan Camilo Benítez Mancipe

2. **Objetivos de la relatoría:** revisar de manera integrada los capítulos 9, 10 y 11 con el fin de identificar vínculos claves entre su contenido y resaltar aspectos esenciales para su desarrollo y presentación. Los capítulos se centraron en tres temas principales: la ética en el diseño y desarrollo de sistemas de IA (Capítulo 9), la gobernanza algorítmica y la toma de decisiones automatizada (Capítulo 10), y el impacto de la IA en el empleo y la economía (Capítulo 11).

3. **Fuentes de información:** los capítulos se basan en diversas fuentes académicas y profesionales, incluyendo:
   - Brey y Dainow (2023) sobre Ethics by Design
   - Pasquale (2015) "La Sociedad de la Caja Negra"
   - Eubanks (2018) "Automating Inequality"
   - O'Neil (2018) "Weapons of Math Destruction"
   - Greene, Hoffmann y Stark (2019) sobre ética en IA/ML
   - Veale y Binns (2017) sobre aprendizaje automático justo

4. **Desarrollo del tema (Síntesis de la relatoría)**
   4.1 Introducción al capítulo o contenido trabajado: se abordaron tres capítulos principales que tratan sobre:

- Principios éticos en el diseño de IA
- Gobernanza algorítmica y toma de decisiones
- Impacto socioeconómico de la IA en el empleo

4.1 Temas centrales abordados: el capítulo 9 introdujo los principios de la ética desde el diseño y destacó la importancia de integrar la ética en la fase de diseño, considerando la transparencia, la explicabilidad, la privacidad y la protección de datos. El capítulo 10 se centró en el control algorítmico y la automatización de la toma de decisiones, destacando la necesidad de legitimidad democrática y mecanismos adecuados de regulación y control. El capítulo 11 examinó el impacto de la IA en el empleo y abordó cuestiones como la automatización, las desigualdades estructurales, la discriminación en los procesos de selección y la ética en los servicios sociales.

4.2 Conclusiones Preliminares

- Se acordó que todos los capítulos deben seguir una estructura uniforme
- Los conceptos deben ser abordados con claridad y ejemplos
- Es necesario un equilibrio entre innovación y ética

5. Desarrollo de la discusión grupal

5.1 Contribuciones relevantes de los participantes: en la discusión grupal se destacó las contribuciones sobre equidad y justicia en el diseño y uso de la IA, así como un análisis crítico del concepto de "caja negra" en la toma de decisiones automatizada. También se debatió el impacto de la automatización sobre el empleo y la economía, y se plantearon propuestas para mejorar la claridad expositiva y estructural de los capítulos.

5.2 Preguntas formuladas: las preguntas fueron: ¿Cómo se puede hacer que el informe sea comprensible para un público no especializado? ¿Es usted responsable de los efectos de la IA? ¿Existe un sistema inteligente que priorice los equipos o la eficiencia? ¿Es cierto que la IA amplifica la desigualdad social?

5.3 Diferentes contrapuntos o interpretaciones: durante la sesión se debatió el equilibrio entre eficiencia técnica y legitimidad democrática, así como la tensión entre la innovación tecnológica y la protección de los derechos fundamentales. También se discutió la posibilidad de automatizar procesos frente a la necesidad de intervención humana continua, así como la distribución de responsabilidades para la implementación y los resultados de los sistemas de IA.

5.4 Comentarios de la docente: la docente ha establecido pautas específicas para mejorar la calidad de los informes. Esto incluye, por ejemplo, definir claramente

los términos utilizados, respaldar el argumento con citas y referencias científicas y discutir diversos enfoques teóricos. También se recomienda escribir párrafos bien estructurados sobre cada principio ético utilizando diferentes autores, incluyendo ejemplos comprensibles y relevantes, utilizando texto que sea comprensible para todos los lectores y asegurando un formato consistente en todos los capítulos para asegurar la coherencia y comprensibilidad del documento final.

Apéndice E. *Protocolo relatoría sesión 5*

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
INGENIERÍA DE SISTEMAS E INFORMÁTICA

SEMINARIO DE INVESTIGACIÓN
ÉTICA E INTELIGENCIA ARTIFICIAL

Sesión: 5
Tema: Capítulo 4. Capacidades tecnológicas y aplicaciones actuales de la IA
Fecha: 12 de marzo de 2025

1. Asistencia: funciones y responsables de la sesión

   Director: Sonia Cristina Gamboa
   Relator: Neyder Fabian Mosquera Niño
   Protocolante: Andrés Felipe Cárdenas Zárate
   Participantes:
   - Yonathan Camilo Benítez Mancipe
   - Juan Pablo Arias Remolina
   - Sonia Marcela Granados Moreno

2. **Objetivos de la relatoría**: explorar las capacidades tecnológicas actuales de la IA y sus aplicaciones en diversas industrias. La relatoría busca proporcionar una comprensión detallada de las tecnologías de IA clave, incluido el procesamiento del lenguaje natural (PLN), la Visión por Computador y el Deep Learning.

3. Fuentes de información:

- Libros académicos como "Speech and language processing" de Jurafsky & Martin (2020)
- Obras clásicas filosóficas como el "Timeo" de Platón y "De Anima" de Aristóteles
- Artículos científicos como el estudio de Esteva et al. (2017) sobre clasificación del cáncer de piel
- Publicaciones especializadas como el trabajo de Chesney & Citron (2019) sobre Deep Fakes
- Documentos técnicos de instituciones como el Departamento de Defensa de Estados Unidos
- Recursos de organizaciones internacionales como la OMS

4. Desarrollo del tema (Síntesis de la relatoría)

4.1 Introducción al capítulo o contenido trabajado: la relatoría introduce la IA como una de las tecnologías más transformadoras del mundo actual, destacando su impacto en la interacción humano-tecnológica y en los procesos de toma de decisiones globales. Se enfatiza la evolución continua de la IA y la necesidad de un desarrollo ético y responsable**.**

4.2 Temas centrales abordados:

Capacidades tecnológicas de la IA:

- Sistemas orientados a objetivos específicos
- Capacidad de análisis en tiempo real
- Autonomía en la toma de decisiones
- Proceso iterativo de desarrollo y mejora

Procesamiento de Lenguaje Natural (PLN):

- Definición y alcance
- Combinación de lingüística computacional y modelos estadísticos
- Aplicaciones como traducción automática y análisis de sentimientos
- Ejemplos de implementación como GPT-4 y asistentes virtuales

Visión por Computador:

- Evolución histórica y fundamentos filosóficos
- Procesamiento de imágenes y reconocimiento de patrones
- Aplicaciones en medicina y diagnóstico
- Implementación en vehículos autónomos y seguridad

Aprendizaje Profundo (*Deep Learning*):

- Diferencias entre aprendizaje humano y computacional
- Funcionamiento de redes neuronales artificiales
- Proceso de entrenamiento y ajuste
- Aplicaciones prácticas

4.3 Conclusiones Preliminares: la IA ha demostrado ser una herramienta transformadora en muchas industrias y ha tenido un impacto significativo en áreas como la salud y la seguridad. Sin embargo, su desarrollo debe equilibrar la eficiencia técnica y las consideraciones éticas, ya que las decisiones automatizadas pueden tener profundas consecuencias para la sociedad. En este contexto, es crucial garantizar la supervisión humana de los sistemas críticos para asegurar la rendición de cuentas, la transparencia y el respeto de los derechos fundamentales.

5. Desarrollo de la discusión grupal

5.1 Contribuciones relevantes de los participantes: Durante el debate surgieron preguntas que nos llevaron a pensar más profundamente sobre el papel de la IA. Por ejemplo, alguien preguntó cómo entrenar los sistemas para valorar la verdad tanto como la rentabilidad. Esto ha provocado un debate sobre los intereses detrás del desarrollo de estas tecnologías. También se discutieron las diferencias entre el aprendizaje asistido por humanos y por computadora, comparando aspectos como la experiencia humana, la intuición y el pensamiento crítico con los procesos estadísticos y algorítmicos de las máquinas. Estas intervenciones ayudaron a enriquecer el debate y a considerar la IA desde una perspectiva más ética y humana**.**

5.2 Preguntas formuladas: las preguntas fueron: ¿Cómo entrenamos sistemas para que valoren la verdad tanto como la rentabilidad? **Y ¿Cómo se diferencia el aprendizaje humano del aprendizaje de una computadora?**

5.3 Diferentes contrapuntos o interpretaciones:

- El uso de la IA en aplicaciones militares y sus implicaciones éticas
- El balance entre beneficios y riesgos de la IA en la difusión de información
- Las diferencias entre el procesamiento humano y artificial de la información

5.4 Comentarios de la docente: la docente apreció la diversidad y profundidad de las fuentes utilizadas, que van desde textos filosóficos clásicos hasta estudios científicos contemporáneos y documentos técnicos, lo que dio solidez al análisis.

También sugirió reforzar el vínculo entre los aspectos técnicos y sus impactos éticos y sociales promoviendo una visión más crítica y global.

Apéndice F. *Protocolo relatoría sesión 6*

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
INGENIERÍA DE SISTEMAS E INFORMÁTICA

SEMINARIO DE INVESTIGACIÓN
ÉTICA E INTELIGENCIA ARTIFICIAL

Sesión: 6
Tema: Capítulo 5. Sesgos, riesgos y desafíos de la IA
Fecha: 19 de marzo de 2025

1. Asistencia: funciones y responsables de la sesión

   Director: Sonia Cristina Gamboa
   Relator: Andrés Felipe Cárdenas Zárate
   Protocolante: Juan Pablo Arias Remolina
   Participantes:
   - Yonathan Camilo Benítez Mancipe
   - Sonia Marcela Granados Moreno
   - Neyder Fabian Mosquera Niño

2. **Objetivos de la relatoría:** discutir los sesgos, riesgos y desafíos asociados a la IA, así como examinar sus implicaciones éticas, sociales, económicas y políticas, con el fin de promover un desarrollo responsable que maximice los beneficios sociales y minimice los daños potenciales**.**

3. Fuentes de información:
   - Libros académicos como "Weapons of Math Destruction" de O'Neil (2016)
   - "Superintelligence: Paths, Dangers, and Strategies" de Bostrom (2014)
   - "The Second Machine Age" de Brynjolfsson y McAfee (2014)
   - Investigaciones académicas como las de Buolamwini y Gebru (2018)
   - Estudios económicos de Frey y Osborne (2017)
   - Trabajos sobre ética y gobernanza como los de Floridi (2019)
   - Análisis sociológicos como "The Age of Surveillance Capitalism" de Zuboff (2019)

4. Desarrollo del tema (Síntesis de la relatoría)

4.1 Introducción al capítulo o contenido trabajado: la relatoría introduce cómo la IA se ha convertido en uno de los desarrollos más disruptivos del siglo XXI, transformando la dinámica social y los límites de las capacidades humanas. Se plantea la necesidad de cuestionar los sesgos implícitos, posibles daños y dilemas que presenta esta tecnología.

4.2 Temas centrales abordados:

Desafíos del sesgo en la IA:

- Sesgo algorítmico y su impacto en la toma de decisiones
- Discriminación automatizada
- Sesgos en datos de entrenamiento
- Impacto en grupos marginados

Amenazas existenciales y consideraciones de seguridad:

- Riesgos de superinteligencia
- Problema de alineación de valores
- Armas autónomas letales
- Seguridad y control humano

Impacto de la automatización en el frente social y laboral:

- Transformación del mercado laboral
- Polarización económica
- Desigualdad y distribución de riqueza
- Desafíos para la identidad y significado social

Vigilancia, autonomía y privacidad:

- Capitalismo de vigilancia
- Amenazas a la privacidad individual
- Sistemas de vigilancia estatal
- Impacto en libertades civiles

Desafíos de gobernanza, regulación y ética:

- Marco regulatorio insuficiente
- Distribución de responsabilidades
- Principios éticos
- Complejidad en la implementación

4.3 Conclusiones Preliminares: la IA tiene el potencial de transformar radicalmente la existencia humana y nos plantea desafíos complejos que requieren soluciones

integrales. En este contexto, es esencial un enfoque proactivo, que moldee el desarrollo de estos sistemas de forma responsable y busque siempre un equilibrio entre innovación y prudencia. Además, la participación democrática en los procesos de gobernanza de la IA cobra cada vez mayor importancia para garantizar que su desarrollo y aplicación sirvan al bien común y no a intereses particulares**.**

5. Desarrollo de la discusión grupal

5.1 Contribuciones relevantes de los participantes: se abordaron las implicaciones del sesgo algorítmico, destacando su impacto en la imparcialidad de las decisiones automatizadas, especialmente en contextos sensibles como el acceso a los servicios públicos, el empleo y la justicia. También se profundizo en las implicaciones económicas de la automatización**.**

5.2 Preguntas formuladas:

- ¿Cómo asegurar un desarrollo responsable de la IA?
- ¿Cómo distribuir equitativamente los beneficios de la automatización?
- ¿Cómo proteger la privacidad y autonomía en la era digital?

5.3 Diferentes contrapuntos o interpretaciones:

- Diferentes enfoques para la gobernanza de la IA
- Distintas perspectivas sobre la responsabilidad algorítmica
- Propuestas para mitigar riesgosa

5.4 Comentarios de la docente: la docente sugirió incluir ejemplos concretos para ilustrar con mayor claridad los riesgos y dilemas asociados al uso de la IA. También recomendó citar a los autores mencionados de forma más explícita para reforzar la conexión entre las fuentes utilizadas y los argumentos presentados. Finalmente, animó a los participantes a examinar críticamente diferentes perspectivas, promoviendo así una reflexión más profunda y consciente sobre el tema.

Apéndice G. *Protocolo relatoría sesión 7*

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
INGENIERÍA DE SISTEMAS E INFORMÁTICA

Universidad Industrial de Santander

SEMINARIO DE INVESTIGACIÓN
ÉTICA E INTELIGENCIA ARTIFICIAL

Sesión: 7
Tema: Capítulo 12. IA, Derechos humanos y sostenibilidad
Fecha: 26 de marzo de 2025

1. Asistencia: funciones y responsables de la sesión

   Director: Sonia Cristina Gamboa
   Relator: Sonia Marcela Granados Moreno
   Protocolante: Yonathan Camilo Benítez Mancipe
   Participantes:
   - Juan Pablo Arias Remolina
   - Andrés Felipe Cárdenas Zárate
   - Neyder Fabian Mosquera Niño

2. **Objetivos de la relatoría:** estudiar la relación entre la IA, los derechos humanos y la sostenibilidad, explorando cómo la IA puede tanto proteger como amenazar los derechos fundamentales, así como su impacto en el desarrollo sostenible.

3. Fuentes de información:
   - Comunicado de la Comisión Europea (2019) sobre los requisitos para una IA fiable
   - Directrices Universales para la Inteligencia Artificial de The Public Voice (2018)
   - Recomendación sobre la Ética de la Inteligencia Artificial de la UNESCO (2021)
   - Documentos de Derechos Digitales sobre IA en América Latina
   - Artículos periodísticos de El País (2025)

4. Desarrollo del tema (Síntesis de la relatoría)

   4.1 Introducción al capítulo o contenido trabajado: Esta relatoría introduce la relación fundamental entre los derechos humanos universales y el desarrollo de la IA, destacando cómo esta tecnología representa tanto oportunidades como amenazas para los derechos fundamentales y la sostenibilidad.

   4.2 Temas centrales abordados:

   Requisitos esenciales para una IA fiable:
   - Intervención y supervisión humanas
   - Solidez y seguridad técnicas
   - Privacidad y gestión de datos

- Transparencia
- Diversidad, no discriminación y equidad
- Bienestar social y medioambiental
- Rendición de cuentas

Protección de derechos humanos en el contexto de la IA:

- Equidad y no discriminación
- Protección de datos personales
- Equidad procesal
- Acceso inclusivo a la tecnología

IA y sostenibilidad

- Optimización de recursos energéticos
- Gestión ambiental
- Impacto ecológico de la IA
- Agricultura de precisión
- Ciudades inteligentes

Desafíos y oportunidades:

- Impacto en el ámbito laboral
- Vigilancia y seguridad pública
- Protección de grupos vulnerables
- Gestión de crisis humanitarias

4.3 Conclusiones Preliminares: Como conclusión preliminar, la IA ofrece un potencial considerable para transformar numerosos aspectos de la sociedad, como el mundo laboral, la vigilancia y la seguridad pública, la protección de grupos vulnerables y la gestión de crisis humanitarias. Estas aplicaciones presentan tanto oportunidades como desafíos, lo que pone de relieve la necesidad de un enfoque ético y responsable en su desarrollo e implementación**.**

5. Desarrollo de la discusión grupal

5.1 Contribuciones relevantes de los participantes: la relatoría abordó aspectos importantes como la necesidad de marcos regulatorios internacionales adecuados, el impacto ambiental de la IA, los desafíos del mercado laboral asociados a la automatización y las especificidades de su implementación en América Latina.

5.2 Preguntas formuladas:

- ¿Cómo garantizar que la IA respete los derechos humanos fundamentales?
- ¿Cómo balancear el desarrollo tecnológico con la sostenibilidad ambiental?
- ¿Qué medidas son necesarias para prevenir la discriminación algorítmica?

- ¿Cómo asegurar una transición tecnológica justa en el ámbito laboral?

5.3 Diferentes contrapuntos o interpretaciones:

- Balance entre innovación tecnológica y protección de derechos
- Tensión entre eficiencia y sostenibilidad
- Debate sobre automatización y derechos laborales
- Perspectivas sobre vigilancia y libertades civiles

5.4 Comentarios de la docente: la docente enfatizó la importancia de un enfoque ético para el desarrollo de la IA, señalando que el progreso tecnológico es insuficiente sin considerar el impacto humano. También enfatizó la necesidad de integrar diferentes perspectivas y contextos en el análisis de esta tecnología, ya que sus impactos varían según el contexto social, cultural y económico. Asimismo, señaló la importancia de marcos regulatorios y mecanismos de supervisión adecuados para el uso responsable de la IA.