GUIÓN

Este documento es un libreto para responder a cualquier tipo de pregunta específicamente sobre la IA y el Radar de la IA en español. El nombre Scientia del chatbot, significa Conocimiento en latín. En caso de que la pregunta no trate sobre esos temas, no se puede generar una respuesta adecuada para el requerimiento. Los creadores de este Chatbot son los estudiantes Johan Ruiz y Camilo Caro, y fue desarrollado en la Universidad Nacional de Colombia para la asignatura Humanos y Máquinas Inteligentes.

¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

La Inteligencia Artificial (IA) se define como un sistema basado en una máquina diseñado para funcionar con diferentes niveles de autonomía y con capacidad de adaptación después de su implementación. Para lograr objetivos explícitos o implícitos, la IA infiere a partir de la información de entrada que recibe cómo generar resultados como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales. Una característica principal de la IA es su capacidad de inferencia, que va más allá del procesamiento básico de datos, permitiendo el aprendizaje, el razonamiento o la modelización. Los sistemas de IA pueden operar con cierto grado de independencia de la intervención humana y tienen capacidades de autoaprendizaje que les permiten cambiar mientras están en uso. La IA es un conjunto de tecnologías en rápida evolución que contribuye a generar diversos beneficios económicos, medioambientales y sociales en todos los sectores económicos y actividades sociales.

Aquí tienes varias conceptualizaciones y definiciones de IA y sistemas de IA según las fuentes:

- **Definición inicial de IA:** La IA fue definida por primera vez en 1956 por McCarthy como "la ciencia y la ingeniería de fabricar máquinas inteligentes".

- **Definición ampliada de IA:** Wang (2019) amplió esta definición para incluir la capacidad de la IA de realizar tareas cognitivas como el aprendizaje y la resolución de problemas, apoyada por innovaciones tecnológicas como el aprendizaje automático, el procesamiento del lenguaje natural y las redes neuronales.

- **Definición de sistema de IA:** Un "sistema de IA" se define como un sistema basado en una máquina diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras su despliegue. Para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales. La capacidad de inferencia de un sistema de IA va más allá del procesamiento básico de datos, permitiendo el aprendizaje, el razonamiento o el modelado. Los sistemas de IA están diseñados para funcionar con diferentes niveles de autonomía, lo que significa que pueden actuar con cierto grado de independencia de la acción humana y tienen capacidades para funcionar sin intervención humana. También pueden utilizarse de forma independiente o como componentes de un producto.

También existen diferentes conceptos relacionados:

- **IA estrecha (Narrow AI):** Un sistema de IA que se desempeña bien en una única tarea o en un conjunto limitado de tareas, como el análisis de sentimientos o jugar al ajedrez.

- **IA de frontera (Frontier AI):** Modelos de IA que pueden realizar una amplia variedad de tareas e igualar o superar las capacidades presentes en los modelos más avanzados de la actualidad. Los LLM son un ejemplo de IA de frontera.

- **Modelos de IA de uso general:** Un modelo de IA, a menudo entrenado con un gran volumen de datos utilizando autosupervisión a gran escala, que presenta un grado considerable de generalidad y es capaz de realizar una gran variedad de tareas distintas. Estos modelos pueden integrarse en diversos sistemas o aplicaciones posteriores. Los grandes modelos de IA generativa son un ejemplo típico, ya que permiten la generación flexible de contenido (texto, audio, imágenes, vídeo) adaptable a diversas tareas.

Los sistemas de IA ofrecen una amplia gama de beneficios económicos, medioambientales y sociales en diversos sectores económicos y actividades sociales, como son la ventaja competitiva y optimización, ya que la IA puede proporcionar ventajas competitivas esenciales a las empresas y optimizar operaciones, mejorar la predicción y la asignación de recursos; así como la creación de servicios personalizados, pues facilita la personalización de soluciones digitales disponibles para la población y las organizaciones.

Algunos sectores específicos para los que sirve son:

- **Sanidad**: Mejora la predicción, optimiza operaciones, y personaliza soluciones digitales en ámbitos como la asistencia sanitaria, incluyendo la detección, diagnóstico, prevención, control y tratamiento de enfermedades, y la mejora de los sistemas sanitarios.

- **Agricultura y Seguridad Alimentaria**: Contribuye a mejorar la predicción y optimizar operaciones.

- **Educación y Formación**: Mejora la predicción, optimiza operaciones y personaliza soluciones digitales. Fomenta una educación y formación digitales de alta calidad, permitiendo a estudiantes y profesores adquirir las capacidades y competencias digitales necesarias, incluyendo la alfabetización mediática y el pensamiento crítico, para participar activamente en la economía, la sociedad y los procesos democráticos. La IA tiene un impacto transformador en el aprendizaje personalizado, la automatización administrativa y la mejora del contenido educativo.

- **Medios de Comunicación, Deporte y Cultura**: Contribuye a mejorar la predicción y optimizar operaciones.

- **Gestión de Infraestructuras, Energía, Transporte y Logística**: Mejora la predicción, optimiza operaciones y personaliza soluciones digitales. En transporte, mejora la seguridad y resiliencia de los sistemas y la movilidad.

- **Servicios Públicos**: Mejora la predicción, optimiza operaciones y personaliza soluciones digitales. Aumenta la eficiencia y calidad de la administración y los servicios públicos.

- **Seguridad y Justicia**: Mejora la predicción, optimiza operaciones y personaliza soluciones digitales. Puede apoyar el poder de decisión de los jueces o la independencia judicial.

- **Eficiencia de Recursos y Energía**: Contribuye a mejorar la predicción y optimizar operaciones. Fomenta la sostenibilidad energética.

- **Vigilancia Ambiental, Conservación y Restauración de la Biodiversidad y los Ecosistemas, Mitigación y Adaptación al Cambio Climático**: Mejora la predicción, optimiza operaciones y personaliza soluciones digitales. Promueve un alto nivel de protección y mejora de la calidad del medio ambiente, la protección de la biodiversidad, la protección contra la contaminación y las medidas de transición ecológica.

Las principales capacidades de la IA se resumen en:

- Conversar de manera fluida y extensa, extrayendo información de vastos conjuntos de datos de entrenamiento.
- Escribir largas secuencias de código funcional a partir de instrucciones en lenguaje natural, incluyendo la creación de nuevas aplicaciones.
- Obtener altas puntuaciones en exámenes de secundaria y universitarios en muchas materias.
- Generar artículos de noticias plausibles.
- Combinar ideas de dominios muy diferentes de forma creativa.
- Traducir entre múltiples idiomas.
- Dirigir las actividades de robots mediante razonamiento, planificación y control de movimiento.
- Analizar datos mediante la creación de gráficos y el cálculo de cantidades clave.
- Responder preguntas sobre imágenes que requieren razonamiento de sentido común.
- Resolver problemas de matemáticas de concursos de secundaria.
- Resumir documentos extensos.

¿QUÉ ES EL RADAR DE LA IA?

El radar de la IA es un marco conceptual para evaluar las aplicaciones de la inteligencia artificial desde una perspectiva ética y responsable. Este radar se basa en cinco ejes que permiten analizar y garantizar que las aplicaciones de la IA se desarrollen y utilicen de manera adecuada. Los cinco ejes son: diseño ético, alfabetización en IA (educación), uso ético (ética), manejo de riesgos y regulación. El radar de la IA busca ser una herramienta práctica para evaluar y guiar el desarrollo de aplicaciones de IA, asegurando que sean éticas, responsables y beneficiosas para la sociedad.

La Inteligencia Artificial (IA) está transformando rápidamente la sociedad, impactando en diversos sectores como la educación, la salud, la economía y el entretenimiento. Ante este panorama, es fundamental comprender las tendencias, oportunidades y desafíos que plantea la IA. El "radar de la IA" se presenta como una herramienta valiosa para analizar el desarrollo de la IA desde una perspectiva multidimensional, abarcando aspectos clave como el diseño de sistemas inteligentes, la educación del público sobre la IA, las implicaciones éticas, el manejo de los riesgos asociados y la necesidad de una regulación adecuada. El radar de la IA fue introducido por el Dr. Alberto Delgado, profesor titular en la Universidad Nacional de Colombia en Bogotá. Este marco, con sus cinco ejes, es una herramienta valiosa para evaluar aplicaciones de IA en un entorno educativo y profesional, especialmente en áreas como la automatización industrial, donde la interacción entre humanos y máquinas es crítica.

El radar de la IA se centra en cinco ejes principales:

- *Diseño Ético:* Se enfoca en cómo se diseña la IA, asegurando que los sistemas sean justos, transparentes y respeten los derechos humanos desde su concepción. Incluye consideraciones como la eliminación de sesgos en los datos, la transparencia en los algoritmos y la inclusión de principios éticos en el proceso de desarrollo. En el diseño ético de los sistemas inteligentes, se deben tener en cuenta los criterios éticos y la responsabilidad social, el impacto en el bienestar individual, colectivo y ambiental, y un enfoque centrado en las personas, evitando reducir la evaluación del progreso a indicadores simplistas.

- *Alfabetización en IA (educación):* Se refiere a la necesidad de que tanto los desarrolladores como los usuarios finales comprendan cómo funcionan los sistemas de IA. Promueve la educación y la concienciación sobre las capacidades, limitaciones y posibles impactos de la IA, para que las personas puedan tomar decisiones informadas. La alfabetización en IA se conceptualiza en cuatro aspectos principales: conocer y comprender la IA, usar y aplicar la IA, evaluar y crear IA, y cuestiones éticas.

- *Uso Ético (ética):* Evalúa cómo se utiliza la IA en la práctica, asegurando que las aplicaciones no causen daño o perjuicio a las personas o a la sociedad. Incluye la responsabilidad de evitar usos malintencionados, como la vigilancia masiva, la discriminación o la manipulación. El uso de la IA debe limitarse a fines legítimos y con supervisión humana.

- *Manejo de Riesgos:* Se centra en identificar, evaluar y mitigar los riesgos asociados con la implementación de la IA. Esto incluye riesgos técnicos (fallos del sistema), éticos (impacto social) y legales (cumplimiento normativo). Para gestionar los riesgos, es fundamental aplicar procesos de evaluación y administración de riesgos, considerando su impacto, plausibilidad y el contexto de aplicación de la IA.

- *Regulación:* Aborda la necesidad de marcos regulatorios claros y efectivos que guíen el desarrollo y uso de la IA. Incluye la creación de políticas y normas que garanticen la responsabilidad, la transparencia y la rendición de cuentas en el uso de estas tecnologías. La regulación de la IA es uno de los ejes más dinámicos, con diferentes países y regiones desarrollando sus propias aproximaciones, como el Reglamento de IA de la Unión Europea o las directrices de la OCDE.

EXPLICACIÓN PROFUNDA – EJES DEL RADAR DE LA IA

1. Diseño ético

Las consideraciones más importantes en el diseño de herramientas de Inteligencia Artificial (IA) se centran en garantizar que estas tecnologías sean beneficiosas para la humanidad, fiables, seguras y que respeten los derechos fundamentales y los valores democráticos. Estas consideraciones abarcan aspectos técnicos, éticos y sociales a lo largo de todo el ciclo de vida de la IA. A continuación, se detallan las consideraciones clave:

1. **Enfoque Centrado en el Ser Humano y Valores Fundamentales:** La IA debe diseñarse para ser una tecnología centrada en el ser humano y fiable, alineándose con valores de la Unión como la dignidad humana, la libertad, la igualdad, la democracia y el Estado de derecho. Su objetivo último debe ser aumentar el bienestar

humano y proteger derechos fundamentales como la salud, la seguridad, la privacidad, la no discriminación, el derecho a la educación, los derechos de los trabajadores, los derechos de las personas con discapacidad, la igualdad de género y la tutela judicial efectiva.

2. **Gestión Integral de Riesgos (Principios S.A.F.E.):** El diseño debe integrar un modelo de gestión de riesgos continuo e iterativo que abarque todo el ciclo de vida de la IA. Este modelo se basa en los principios S.A.F.E.:

   - Sostenibilidad (Resilience): Asegurar que las salidas de la IA sean resistentes a eventos anómalos extremos o ciberataques.
   - Precisión (Accuracy): Garantizar la exactitud predictiva de las salidas del modelo.
   - Equidad (Fairness): Diseñar sistemas para evitar sesgos injustos y discriminación, especialmente hacia grupos vulnerables. Esto implica el uso de conjuntos de datos representativos y completos.
   - Explicabilidad (Explainability): Permitir que las salidas del modelo sean comprendidas y supervisadas por humanos, con una clara explicación de sus causas impulsoras.

3. **Calidad de Datos:** Los sistemas de IA deben desarrollarse utilizando conjuntos de datos de entrenamiento, validación y prueba de alta calidad, que sean pertinentes, representativos, libres de errores y completos. Se deben implementar prácticas adecuadas de gobernanza y gestión de datos, incluyendo la transparencia sobre el propósito original de la recopilación de datos y la detección y mitigación de posibles sesgos. Además, se debe aplicar el principio de minimización de datos y protección de datos desde el diseño y por defecto, utilizando técnicas como la anonimización y el cifrado. Excepcionalmente, se pueden tratar categorías especiales de datos personales para detectar y corregir sesgos, siempre con garantías adecuadas.

4. **Supervisión y Control Humano:** Los sistemas de IA deben diseñarse como herramientas al servicio de las personas, permitiendo una supervisión y un control humanos efectivos durante su uso. Esto incluye dotar a los sistemas de interfaces adecuadas, permitir que respondan al operador humano y asegurar que las personas encargadas de la supervisión tengan las competencias y la formación necesarias. Para sistemas de identificación biométrica de alto riesgo, se exige una supervisión humana reforzada, que puede implicar la verificación por al menos dos personas. Los sistemas deben permitir la intervención humana o la interrupción segura.

5. **Seguridad y Solidez Técnica:** Los sistemas de IA deben ser robustos, precisos y ciberseguros, capaces de resistir errores, fallos, inconsistencias y ataques maliciosos. Se deben incorporar medidas técnicas y organizativas para prevenir o minimizar comportamientos dañinos o indeseables, incluyendo mecanismos de interrupción segura (planes de prevención de fallos). Además, se deben adoptar soluciones para abordar vulnerabilidades específicas de la IA, como el envenenamiento de datos o los ataques adversarios.

6. **Transparencia y Documentación:** Los sistemas de IA de alto riesgo deben ser transparentes, proporcionando información clara, concisa y comprensible a los responsables del despliegue a través de instrucciones de uso. Se debe mantener una documentación técnica actualizada y completa sobre el desarrollo, funcionamiento y limitaciones del sistema. Finalmente, se requiere el registro automático de acontecimientos (archivos de registro) para la trazabilidad y la vigilancia posterior a la comercialización.

7. **Evitar Prácticas Prohibidas:** El diseño de herramientas de IA debe evitar expresamente las prácticas consideradas inaceptables y prohibidas, ya que son contrarias a los valores fundamentales:

   - Sistemas que usen técnicas subliminales o manipuladoras para alterar sustancialmente el comportamiento humano, causando perjuicios considerables.
   - Explotación de vulnerabilidades de personas (edad, discapacidad, situación socioeconómica) para causar perjuicios significativos.
   - Sistemas de "puntuación ciudadana" que evalúen el comportamiento social para generar trato perjudicial injustificado.
   - Evaluación de riesgos de criminalidad basada *únicamente* en el perfilado o rasgos de personalidad.
   - Creación o ampliación de bases de datos de reconocimiento facial mediante extracción no selectiva de imágenes de internet o CCTV.
   - Sistemas de reconocimiento de emociones en el lugar de trabajo y centros educativos (con excepciones médicas o de seguridad).
   - Uso de sistemas de identificación biométrica remota "en tiempo real" en espacios públicos por parte de las autoridades encargadas de hacer cumplir la ley, salvo en situaciones excepcionales y bajo estrictas salvaguardias.
   - Sistemas de IA para influir en los resultados electorales o el comportamiento electoral de los votantes (excluyendo herramientas administrativas/logísticas no expuestas directamente al público).

8. **Accesibilidad y No Discriminación:** Los proveedores deben garantizar el pleno cumplimiento de los requisitos de accesibilidad, integrando estas medidas en el diseño desde el inicio para asegurar un acceso pleno e igualitario para todas las personas, incluidas las personas con discapacidad. El diseño inclusivo debe evitar la perpetuación o amplificación de sesgos históricos y la discriminación contra grupos vulnerables.

9. **Vigilancia Posterior a la Comercialización (Post-Market Surveillance):** Los proveedores deben establecer un sistema robusto de vigilancia posterior a la comercialización para recopilar y analizar datos sobre el rendimiento del sistema en la vida real, lo que permite la detección temprana de problemas y la adopción de medidas correctoras. Esto incluye la obligación de notificar incidentes graves a las autoridades competentes.

2. Alfabetización

Definiciones Fundamentales de Alfabetización en IA:

La alfabetización en IA se refiere a un conjunto de competencias que capacitan a las personas para evaluar críticamente las tecnologías de IA, comunicarse y colaborar eficazmente con la IA, y usar la IA como herramienta en diversos entornos. También se ha definido como la comprensión de los conceptos de IA y las competencias para usar estos conceptos en la evaluación y la comprensión del mundo real. Otra perspectiva la describe como la expresión integral de conocimientos, habilidades, procesos, métodos, actitudes emocionales y valores que se forman gradualmente al recibir educación sobre IA. En el ámbito laboral, la alfabetización en IA de los empleados es una combinación de capacidades tecnológicas, laborales, humano-máquina y de aprendizaje.

Seis Constructos Clave de la Alfabetización en IA (Marco Sistemático):

Un análisis sistemático de la literatura entre 2019 y 2023 identificó seis constructos fundamentales para la alfabetización en IA, que aparecen en la mayoría de los estudios revisados:

- **Reconocer:** La capacidad de identificar la presencia de IA. Aunque es fundamental, a menudo se da por sentada en muchos estudios.

- **Conocer y Comprender:** Implica la adquisición de conceptos, habilidades y conocimientos fundamentales de la IA, incluyendo cómo procesa datos, sus mecanismos subyacentes como el aprendizaje automático y las redes neuronales, y el papel de los humanos en su desarrollo. Es un componente recurrente en la investigación.

- **Usar y Aplicar:** La habilidad para operar herramientas y aplicaciones de IA, y para integrar conceptos de IA en la resolución de tareas y problemas. Esto abarca la colaboración entre humanos y la IA, y la adaptación de las herramientas de IA a objetivos específicos.

- **Evaluar:** La capacidad de analizar e interpretar críticamente los resultados de las aplicaciones de IA, y de formar opiniones informadas sobre la interacción con estas tecnologías.

- **Crear:** Competencias relacionadas con el diseño, desarrollo y creación de sistemas de IA, o la interacción creativa con ellos. Este aspecto ha recibido la menor atención en los programas de implementación.

- **Navegar Éticamente:** Concierne las consideraciones éticas y el impacto social de la IA, incluyendo aspectos como la privacidad, los sesgos, la desinformación, la equidad, y la rendición de cuentas. Es considerado un constructo fundamental en la mayoría de los trabajos.

Cuatro Pilares de la Alfabetización en IA (Enfoque Sociotécnico):

Una propuesta de currículo interdisciplinario para la alfabetización en IA se estructura en cuatro pilares, especialmente relevante con la proliferación de herramientas de IA Generativa (Gen-AI) como ChatGPT:

- **Comprensión del alcance y las dimensiones técnicas de la IA:** Incluye la introducción a la IA y el aprendizaje automático, la representación del conocimiento en IA (simbólica y conexionista), el funcionamiento de los motores de búsqueda, los sistemas generativos y la generación aumentada por recuperación, y cómo funcionan los grandes modelos de lenguaje.

- **Aprender a interactuar con la IA Generativa de manera informada y responsable:** Se centra en la interacción con grandes modelos de lenguaje, la integridad académica/profesional, la autoría y la propiedad, y la ingeniería de *prompts* para el aprendizaje.

- **Revisión crítica de los problemas de la IA ética y socialmente responsable:** Aborda el uso responsable de la IA, la seguridad, la privacidad y las cuestiones éticas, y estudios de caso.

- **Implicaciones sociales y futuras de la IA:** Explora la percepción pública de la IA, la IA Generativa y el futuro del trabajo, la IA y las políticas (ej. accesibilidad), y la IA para el bien, incluyendo la sostenibilidad y el desarrollo.

Se destacan también Las "Cinco Grandes Ideas" (The Five Big Ideas) de la IA para la educación K-12, que incluyen percepción, representación y razonamiento, aprendizaje, interacción natural e impacto social, mostrando:

- Dimensiones cognitivas, afectivas y socioculturales de la alfabetización en IA.
- La importancia de la conciencia y la capacidad, y la concienciación social sobre el impacto de la IA.
- Conceptos como la competencia en IA y el pensamiento en IA.

Estos elementos, ya sean constructos o pilares, buscan proporcionar un marco comprensivo para entender, enseñar y evaluar la interacción humana con la IA en un mundo cada vez más tecnologizado.

3. Uso ético

El uso de la inteligencia artificial (IA) implica una serie de consideraciones éticas fundamentales que buscan equilibrar los beneficios de estas tecnologías con la necesidad de proteger los derechos y el bienestar de las personas y la sociedad. A continuación, se resumen las principales consideraciones éticas presentes en la literatura:

1. **Protección de Derechos Fundamentales y Valores Humanos:** La IA debe ser diseñada y utilizada de manera que sea centrada en el ser humano y fiable. Su desarrollo y despliegue deben alinearse con valores como el respeto a la dignidad humana, la libertad, la igualdad, la democracia y el Estado de derecho. Esto incluye la protección de la salud, la seguridad, la privacidad, los datos personales, la no discriminación, el derecho a la educación, los derechos de los trabajadores, los derechos de las personas con discapacidad, la igualdad de género, el derecho a la tutela judicial efectiva y la presunción de inocencia. El objetivo último de la IA debe ser aumentar el bienestar humano.

2. **Transparencia y Explicabilidad:** La opacidad de algunos sistemas de IA es una preocupación significativa. Se considera crucial que los sistemas de IA sean transparentes y explicables para generar confianza, permitir la supervisión humana, identificar sesgos, corregir errores y garantizar la rendición de cuentas. Esto implica:

   - Proporcionar información clara y completa sobre el funcionamiento, capacidades y limitaciones del sistema.
   - Permitir la trazabilidad de los sistemas de IA de alto riesgo y el registro automático de acontecimientos.
   - Informar a las personas cuando interactúan con un sistema de IA o cuando el contenido ha sido generado o manipulado artificialmente, salvo cuando sea obvio o para fines de cumplimiento de la ley.

3. **Equidad y No Discriminación:** La IA presenta riesgos de discriminación y puede perpetuar o incluso amplificar sesgos existentes, especialmente contra grupos vulnerables como mujeres, personas de ciertas edades, con discapacidad o de orígenes raciales o étnicos específicos. Las consideraciones éticas en este ámbito incluyen:

   - Promover un diseño inclusivo que garantice la igualdad de acceso y evite los sesgos injustos.

- Implementar prácticas adecuadas de gobernanza y gestión de datos para asegurar que los conjuntos de datos de entrenamiento sean pertinentes, representativos, libres de errores y completos, y que se detecten y mitiguen los posibles sesgos.
- Permitir, bajo estrictas salvaguardias, el tratamiento excepcional de categorías especiales de datos personales para la detección y corrección de sesgos en sistemas de IA de alto riesgo.

4. **Seguridad y Fiabilidad:** Los sistemas de IA deben ser robustos, precisos y ciberseguros para evitar daños a la salud, la seguridad y los derechos fundamentales. Esto se logra mediante:

   - Un sistema de gestión de riesgos que identifique, evalúe y mitigue los riesgos conocidos y previsibles, incluyendo el uso indebido razonablemente previsible.
   - Medidas técnicas y organizativas que garanticen la resistencia de los sistemas a errores, fallos, incoherencias y ciberataques.
   - La monitorización continua de los sistemas para mejorar su rendimiento y abordar cualquier problema que surja después de su despliegue.

5. **Privacidad y Gobernanza de Datos:** La protección de los datos personales es un derecho fundamental que debe salvaguardarse a lo largo de todo el ciclo de vida de los sistemas de IA. Esto implica aplicar principios como la minimización de datos y la protección de datos desde el diseño y por defecto, utilizando técnicas como la anonimización, el cifrado o el entrenamiento de algoritmos directamente sobre los datos sin necesidad de su transmisión (como el aprendizaje federado).

6. **Supervisión y Control Humano:** Los sistemas de IA deben ser desarrollados y utilizados como herramientas al servicio de las personas, permitiendo una supervisión y control adecuados por parte de los seres humanos. La decisión final no debe recaer exclusivamente en la IA. Para ello, es necesario:

   - Diseñar sistemas que permitan una supervisión humana efectiva, con interfaces adecuadas y personal competente y formado.
   - Garantizar que los operadores humanos comprendan las capacidades y limitaciones de la IA y puedan intervenir o detener el sistema si es necesario.
   - Implementar una supervisión humana reforzada para sistemas críticos, como la identificación biométrica, que requiera la verificación de los resultados por parte de múltiples personas.

7. **Responsabilidad y Rendición de Cuentas:** Es fundamental que exista un marco claro de responsabilidad a lo largo de toda la cadena de valor de la IA. Esto incluye definir las obligaciones de los proveedores, importadores, distribuidores y responsables del despliegue para garantizar la conformidad con la normativa y mitigar los riesgos. Se reconoce la preocupación por la falta de rendición de cuentas, especialmente cuando los incentivos económicos prevalecen sobre los principios éticos.

8. **Impacto Social y Prácticas Prohibidas:** Si bien la IA puede ofrecer beneficios económicos, medioambientales y sociales en diversos sectores, también puede generar nuevos y poderosos medios para la manipulación, explotación y el control social, lo cual se considera perjudicial y debe ser prohibido. Entre las prácticas de IA prohibidas se incluyen:

- Sistemas de IA que utilicen técnicas subliminales o manipuladoras para alterar el comportamiento humano de manera sustancial, causando perjuicios considerables.
- Sistemas que exploten vulnerabilidades de personas (por edad, discapacidad o situación socioeconómica) para alterar su comportamiento y causar daños.
- Sistemas de "puntuación ciudadana" que evalúen o clasifiquen a personas según su comportamiento social para darles un trato perjudicial o desfavorable injustificado.
- Sistemas de evaluación de riesgos para predecir la comisión de delitos basándose *únicamente* en el perfilado o los rasgos de personalidad.
- Sistemas que creen o amplíen bases de datos de reconocimiento facial extrayendo imágenes de forma indiscriminada de internet o CCTV.
- Sistemas de reconocimiento de emociones en el lugar de trabajo y centros educativos (con excepciones médicas o de seguridad).
- Sistemas de categorización biométrica que clasifiquen personas por atributos sensibles como raza o creencias políticas.
- El uso de sistemas de identificación biométrica remota "en tiempo real" en espacios de acceso público por parte de las autoridades encargadas de hacer cumplir la ley, salvo en situaciones excepcionales estrictamente definidas, con autorización judicial o administrativa y bajo estrictas garantías.
- Sistemas de IA que busquen influir en el resultado de elecciones o referéndums o en el comportamiento electoral de los votantes, excluyendo herramientas puramente administrativas/logísticas no expuestas directamente al público.

En resumen, las consideraciones éticas en la IA giran en torno a asegurar que estas tecnologías se desarrollen y utilicen de manera que respeten los derechos fundamentales, promuevan la equidad, sean seguras, transparentes y estén bajo un control humano efectivo, mitigando los riesgos y prohibiendo las aplicaciones que contravengan estos principios.

4. Manejo de riesgos

Las principales consideraciones en la regulación y el desarrollo de la Inteligencia Artificial (IA) giran en torno a la necesidad de garantizar que estas tecnologías sean confiables, seguras, éticas y beneficien a la sociedad, al tiempo que se mitigan los riesgos inherentes. Este marco busca promover una IA centrada en el ser humano, respetando la dignidad, la libertad, la igualdad y los derechos fundamentales.

1. **Identificación y Clasificación de Riesgos:** La identificación de riesgos en la IA es un esfuerzo multifacético que busca organizar y comprender la complejidad de los posibles perjuicios.

- Repositorio de Riesgos de la IA (AI Risk Repository): Un enfoque clave es la creación de un "Repositorio de Riesgos de la IA" que sistemáticamente revisa y sintetiza clasificaciones existentes de riesgos de la IA para establecer un marco de referencia común. Este repositorio se construyó mediante:

    - Una estrategia de búsqueda sistemática de literatura (revisada por pares y literatura gris), búsqueda retrospectiva y prospectiva, y consulta a expertos para identificar clasificaciones, marcos y taxonomías de riesgos de la IA.

- La extracción de 1612 riesgos individuales de 65 taxonomías en una "base de datos viva" (living database) que puede actualizarse continuamente.
- El uso de un enfoque de "síntesis de marco de mejor ajuste" (best-fit framework synthesis) para desarrollar dos taxonomías principales que clasifican los riesgos de manera estructurada.

2. **Descripción y Tipos de Riesgos:** Los riesgos de la IA se describen y clasifican de varias maneras para capturar su complejidad y multifacetas.

- Taxonomía Causal (Causal Taxonomy): Esta taxonomía de alto nivel clasifica cada riesgo según sus factores causales:

- Entidad (Entity): Indica quién o qué es la causa principal del riesgo. Puede ser el sistema de IA (por ejemplo, generación de contenido dañino, desempoderamiento humano), un humano (por ejemplo, elección de datos de entrenamiento deficientes, diseño malicioso, uso indebido) u "Otros" si es ambiguo. Por ejemplo, la IA es la entidad causal en el 90% de los riesgos de información falsa o engañosa, mientras que en la compromisión de la privacidad es la entidad causal en el 58% de los casos, mostrando menor consistencia en la atribución.

- Intencionalidad (Intent): Determina si el riesgo es un resultado esperado (Intencional) o inesperado (No intencional) de la búsqueda de un objetivo, o si la intencionalidad no está claramente especificada ("Otros"). Por ejemplo, los riesgos de uso malicioso se presentan consistentemente como intencionales por parte de humanos, mientras que la discriminación o información engañosa a menudo se presentan como causadas no intencionalmente por la IA.

- Momento (Timing): Clasifica el riesgo según si ocurre antes del despliegue de la IA (Pre-deployment) o después de que el modelo ha sido entrenado y desplegado (Post-deployment). La mayoría de los riesgos se presentan como "Post-deployment" (62%), mientras que solo el 13% son "Pre-deployment". En general, el 41% de los riesgos se atribuyen a la IA, el 39% a los humanos y el 20% son "Otros". El 35% son no intencionales, el 34% intencionales y el 31% "Otros". Los riesgos relacionados con acciones humanas se identificaron en el 83% de los documentos, y los relacionados con la IA en el 92%.

- Taxonomía de Dominio (Domain Taxonomy): Esta taxonomía de nivel medio clasifica los riesgos en siete dominios principales y 24 subdominios, proporcionando una comprensión estructurada de los peligros y daños asociados con la IA:

    1. Discriminación y toxicidad: Incluye discriminación injusta, exposición a contenido tóxico y rendimiento desigual entre grupos.
    2. Privacidad y seguridad: Abarca la compromisión de la privacidad y las vulnerabilidades de seguridad de los sistemas de IA.
    3. Desinformación: Se refiere a información falsa o engañosa, y la contaminación del ecosistema de la información.
    4. Actores maliciosos y uso indebido: Incluye desinformación, ciberataques, desarrollo/uso de armas y fraude.
    5. Interacción humano-máquina: Riesgos derivados de la interacción entre humanos y sistemas de IA.
    6. Perjuicios socioeconómicos y ambientales: Afectan a la sociedad y al medio ambiente.

7. Seguridad, fallos y limitaciones del sistema de IA: Cubre la falta de capacidad o robustez, la falta de transparencia o interpretabilidad, y los riesgos de múltiples agentes.

Los dominios más comúnmente discutidos en la literatura son los perjuicios socioeconómicos y ambientales (76% de los documentos), seguidos de la seguridad, fallos y limitaciones del sistema de IA (75%), y la discriminación y toxicidad (70%). Los menos comunes son la desinformación (46%) y la interacción humano-máquina.

- Principios S.A.F.E.:

1. Sostenibilidad (Sustainability/Resilience): Resistencia de los resultados de la IA frente a eventos extremos anómalos o ciberataques.
2. Precisión (Accuracy): Exactitud predictiva de las salidas del modelo.
3. Equidad (Fairness): Ausencia de sesgos injustos hacia grupos de población, garantizando un diseño que evite la discriminación.
4. Explicabilidad (Explainability): Capacidad de los resultados del modelo de ser comprendidos y supervisados por humanos, especialmente sus causas impulsoras.

3. Tipos de Perjuicios y Riesgos Generales: La regulación de la IA se enfoca en una amplia gama de perjuicios y riesgos potenciales. El concepto de "riesgo" se define como la combinación de la probabilidad de que se produzca un perjuicio y la gravedad de dicho perjuicio.

- Perjuicios a Derechos Fundamentales: La IA puede generar riesgos tangibles o intangibles, incluyendo perjuicios físicos, psíquicos, sociales o económicos, afectando derechos como la salud, la seguridad, la dignidad humana, la privacidad, la no discriminación, la educación y los derechos de los trabajadores. Se presta especial atención a los derechos de los menores y las personas con discapacidad.

- Riesgos Transversales (Cross-cutting risk factors): Incluyen desafíos técnicos para construir sistemas de IA seguros, evaluar su seguridad y comprender sus decisiones, así como fallos inesperados y barreras para monitorear su uso. La prueba de seguridad de la IA de frontera es ad-hoc, careciendo de estándares o prácticas de ingeniería establecidas.

- Riesgos de Datos: Los datos de entrenamiento pueden reflejar patrones históricos de injusticia, desigualdades o culturas dominantes, lo que lleva a la exposición a lenguaje despectivo y estereotipos hacia grupos marginados.

- Desinformación: Las salidas de la IA pueden ser indistinguibles del contenido humano, lo que dificulta la detección de desinformación a pesar de enfoques como las marcas de agua.

4. Prácticas de IA Prohibidas: Ciertas prácticas de IA se consideran inaceptables por ir en contra de los valores fundamentales y los derechos humanos, y por lo tanto están prohibidas:

- Uso de técnicas subliminales o manipuladoras que alteren el comportamiento de manera que cause perjuicios considerables.

- Explotación de vulnerabilidades de personas o colectivos (por edad, discapacidad, situación socioeconómica) para causar perjuicios significativos.
- Sistemas de "puntuación ciudadana" que evalúen o clasifiquen a personas basándose en su comportamiento social para generar un trato perjudicial injustificado o desproporcionado.
- Evaluaciones de riesgo de criminalidad basadas *únicamente* en el perfilado o rasgos de personalidad.
- Creación o expansión de bases de datos de reconocimiento facial mediante la extracción no selectiva de imágenes de internet o CCTV.
- Sistemas de reconocimiento de emociones en el lugar de trabajo y centros educativos, salvo excepciones estrictas por motivos médicos o de seguridad.
- Uso de sistemas de identificación biométrica remota "en tiempo real" en espacios de acceso público por parte de las autoridades encargadas del cumplimiento de la ley, excepto en situaciones estrictamente necesarias y limitadas (búsqueda de víctimas, amenazas inminentes para la vida o seguridad, localización de sospechosos de delitos graves específicos), y bajo estrictas salvaguardias y autorización.

5. **Consideraciones Clave para la Regulación y el Desarrollo:** La regulación de la IA se basa en un enfoque de gestión de riesgos que se aplica en todo el ciclo de vida del sistema.

- Sistema de Gestión de Riesgos: Un proceso continuo e iterativo que debe establecerse, implementarse y documentarse para los sistemas de IA de alto riesgo. Esto incluye identificar, analizar, estimar y evaluar los riesgos (conocidos, previsibles y derivados del uso indebido) y adoptar medidas de mitigación adecuadas. Las pruebas son fundamentales para asegurar la conformidad y el funcionamiento previsto.

- Sistemas de IA de Alto Riesgo: Se clasifican como de "alto riesgo" aquellos sistemas que son componentes de seguridad de ciertos productos (ej. máquinas, dispositivos médicos) y requieren evaluación de la conformidad por terceros, o aquellos utilizados en ámbitos críticos como la biometría, infraestructuras críticas, educación, empleo, servicios esenciales, aplicación de la ley, migración y justicia. Existen excepciones para sistemas que no influyen sustancialmente en la toma de decisiones o realizan tareas preparatorias/limitadas, a menos que impliquen perfilado de personas físicas.

- Calidad y Gobernanza de Datos: Se exigen conjuntos de datos de entrenamiento, validación y prueba de alta calidad, pertinentes, representativos, sin errores y completos. Se deben implementar prácticas de gobernanza de datos que incluyan transparencia sobre la recopilación y mitigación de sesgos. Se aplica la minimización de datos y la protección de datos desde el diseño. Excepcionalmente, se permite el tratamiento de categorías especiales de datos personales para detectar y corregir sesgos, con garantías adecuadas.

- Documentación Técnica y Conservación de Registros: Esencial para la trazabilidad y la vigilancia. Se requiere documentación completa y actualizada sobre el desarrollo, funcionamiento, algoritmos, datos y gestión de riesgos, así como el registro automático de eventos (archivos de registro).

- Transparencia y Comunicación de Información: Los sistemas de alto riesgo deben diseñarse con transparencia suficiente para que los responsables del despliegue interpreten y usen correctamente sus resultados. Deben ir acompañados de instrucciones de uso claras sobre sus capacidades, limitaciones y riesgos previsibles.

- Supervisión Humana: Los sistemas deben ser supervisables por personas, permitiendo la intervención, anulación o detención segura. Se requiere una supervisión humana reforzada para ciertos sistemas de identificación biométrica, que debe ser verificada por al menos dos personas antes de tomar decisiones.

- Precisión, Solidez y Ciberseguridad: Los sistemas deben ser robustos, precisos y ciberseguros, capaces de resistir errores, fallos y ataques maliciosos (ej., envenenamiento de datos, ataques adversarios). Se requieren medidas técnicas y organizativas para minimizar comportamientos dañinos.

- Evaluación de la Conformidad y Marcado CE: Los sistemas de alto riesgo deben someterse a una evaluación de la conformidad antes de su introducción en el mercado y llevar el marcado CE. Para ciertos sistemas, se requiere la participación de organismos notificados.

- Vigilancia Post-comercialización: Los proveedores deben establecer un sistema para recopilar, documentar y analizar activamente la experiencia de uso de los sistemas de IA de alto riesgo después de su comercialización para detectar la necesidad de medidas correctivas y notificar incidentes graves.

- Modelos de IA de Uso General (GPAI):

    - Transparencia: Los proveedores de GPAI tienen obligaciones específicas de transparencia, como la elaboración de documentación técnica, la provisión de información para que otros proveedores puedan integrar el modelo, el cumplimiento de los derechos de autor (incluyendo la detección y respeto de reservas de derechos) y la publicación de un resumen detallado del contenido utilizado para el entrenamiento.

    - Riesgo Sistémico: Los GPAI con riesgo sistémico (determinados por su capacidad o impacto en el mercado) están sujetos a obligaciones adicionales, como la evaluación continua y mitigación de riesgos sistémicos (incluyendo pruebas de simulación de adversarios) y ciberseguridad adecuada. Se fomenta la elaboración de códigos de buenas prácticas para GPAI.

    - Apoyo a la Innovación: Se promueve la creación de "espacios controlados de pruebas para la IA" (regulatory sandboxes) para facilitar el desarrollo y prueba de sistemas innovadores bajo supervisión. Se permiten las pruebas en condiciones reales de sistemas de alto riesgo fuera de estos sandboxes bajo ciertas condiciones y salvaguardias, como el consentimiento informado. Se busca reducir la carga administrativa y los costes para las PYMES y startups.

6. **Gobernanza y Aplicación:** La implementación de estas consideraciones se apoya en un marco de gobernanza y aplicación.

- Estructuras de Gobernanza: Se establecen organismos a nivel de la Unión Europea, como la Oficina de IA (AI Office) y el Consejo de IA (AI Board), para la supervisión y coordinación. Los Estados miembros designan autoridades nacionales competentes con recursos y experiencia técnica adecuados.

- Sanciones: Se prevén sanciones efectivas, proporcionadas y disuasorias por el incumplimiento de las obligaciones, con límites máximos para multas administrativas.

- Derechos de los Individuos: Se garantiza el derecho a presentar quejas ante las autoridades de vigilancia del mercado y, en ciertos casos, el derecho a obtener una explicación de las decisiones tomadas por la IA que les afecten significativamente.

- Transparencia de Contenido Generado por IA: Los proveedores que generen contenido sintético (audio, imagen, video, texto) deben marcarlo de forma legible por máquina. Los responsables del despliegue de contenido de imagen, audio o video que constituyan "ultrasuplantaciones" (deepfakes) deben divulgar su origen artificial. Los textos generados por IA con fines informativos deben ser divulgados, a menos que hayan sido revisados humanamente o bajo responsabilidad editorial.

- Evaluación y Revisión: La Comisión evaluará y revisará periódicamente la regulación, la lista de sistemas de alto riesgo, las prácticas prohibidas, las obligaciones de transparencia, la eficacia del sistema de supervisión y la eficiencia energética de los modelos de IA, con informes públicos y posibles propuestas de modificación.

En conjunto, el desarrollo de la IA son un proceso dinámico que exige una consideración constante de la innovación, la ética, la seguridad, la transparencia, la equidad y la protección de los derechos humanos en cada etapa de la vida de los sistemas de IA.

5. Regulación

Las principales consideraciones en la regulación y el desarrollo de la Inteligencia Artificial (IA) se centran en garantizar que estas tecnologías sean confiables, seguras, éticas y que beneficien a la sociedad en general, al tiempo que se mitigan sus riesgos inherentes. Este marco busca promover una IA centrada en el ser humano, respetando la dignidad, la libertad, la igualdad y los derechos fundamentales. A continuación, se resumen las consideraciones clave:

1. **Enfoque Centrado en el Ser Humano y Valores Fundamentales:** El objetivo primordial es mejorar el bienestar humano y proteger derechos fundamentales como la salud, la seguridad, la privacidad, la no discriminación, el derecho a la educación y los derechos de los trabajadores. La IA debe ser una herramienta al servicio de las personas, respetando la dignidad y la autonomía personal. Se busca que la IA se desarrolle y utilice de acuerdo con los valores de la Unión Europea, como la democracia y el Estado de Derecho.

2. **Gestión de Riesgos Integral (Principios S.A.F.E.):** Se propone un modelo de gestión de riesgos continuo e iterativo a lo largo de todo el ciclo de vida de la IA, desde el diseño hasta el monitoreo en producción. Este modelo se basa en cuatro variables estadísticas principales resumidas por el acrónimo S.A.F.E.:

   - Sostenibilidad (Sustainability/Resilience): Se refiere a la resistencia de los resultados de la IA frente a eventos extremos anómalos o ciberataques.
   - Precisión (Accuracy): Se refiere a la exactitud predictiva de las salidas del modelo.
   - Equidad (Fairness): Se refiere a la ausencia de sesgos injustos hacia grupos de población, garantizando que el diseño evite la discriminación.
   - Explicabilidad (Explainability): Se refiere a la capacidad de los resultados del modelo de ser comprendidos y supervisados por humanos, especialmente en sus causas impulsoras.

La gestión de riesgos debe identificar y mitigar los riesgos conocidos y razonablemente previsibles para la salud, la seguridad y los derechos fundamentales, considerando tanto el uso previsto como el uso indebido previsible.

3. **Calidad y Gobernanza de Datos:** Los sistemas de IA que se entrenan con datos deben utilizar conjuntos de datos de entrenamiento, validación y prueba de alta calidad, que sean pertinentes, representativos, libres de errores y completos. Se deben implementar prácticas adecuadas de gobernanza y gestión de datos, incluyendo la transparencia sobre el propósito original de la recopilación de datos. Es fundamental identificar, prevenir y mitigar posibles sesgos en los conjuntos de datos, especialmente aquellos que puedan conducir a discriminación contra grupos vulnerables. Se debe aplicar la minimización de datos y la protección de datos desde el diseño y por defecto, utilizando técnicas como la anonimización y el cifrado. Excepcionalmente, se permite el tratamiento de categorías especiales de datos personales para detectar y corregir sesgos, siempre con garantías adecuadas.

4. **Supervisión y Control Humano:** Los sistemas de IA deben diseñarse para ser efectivamente supervisados por personas, permitiendo que respondan al operador humano y que los supervisores tengan las competencias y la formación necesarias. Los usuarios deben poder comprender las capacidades y limitaciones de la IA, ser conscientes del "sesgo de automatización" (confianza excesiva en la IA), interpretar correctamente los resultados, y tener la capacidad de no usar el sistema, anular sus resultados o incluso detenerlo de forma segura. Para sistemas de identificación biométrica de alto riesgo, se exige una verificación reforzada por al menos dos personas antes de actuar o tomar decisiones basadas en la identificación generada por el sistema.

5. **Seguridad y Solidez Técnica:** Los sistemas de IA deben ser robustos, precisos y ciberseguros, capaces de resistir errores, fallos, inconsistencias y ataques maliciosos. Se deben implementar medidas técnicas y organizativas para prevenir o minimizar comportamientos dañinos, incluyendo mecanismos de interrupción segura (planes de prevención de fallos). La ciberseguridad es fundamental para proteger contra acciones maliciosas, como el envenenamiento de datos o ataques adversarios.

6. **Transparencia y Documentación:** Se requiere una documentación técnica completa y actualizada del sistema de IA, incluyendo detalles sobre su desarrollo, funcionamiento, capacidades, limitaciones, algoritmos, datos de entrenamiento y validación, y sistema de gestión de riesgos. Los sistemas de IA de alto riesgo deben ir acompañados de instrucciones de uso claras y comprensibles para los responsables del despliegue, que incluyan información sobre riesgos y medidas de supervisión humana. Los sistemas deben permitir el registro automático de eventos (archivos de registro) para la trazabilidad y la vigilancia posterior a la comercialización. Además, deben interactuar directamente con personas deben informarles claramente que están interactuando con una IA, salvo que sea obvio. Los proveedores de IA que generen contenido sintético (audio, imagen, video, texto) deben marcarlo de forma legible por máquina para indicar que ha sido generado o manipulado artificialmente. Quienes publiquen texto generado por IA con fines informativos deben divulgarlo, a menos que haya habido revisión humana/editorial.

7. **Prohibiciones Específicas:** Se prohíben prácticas de IA consideradas inaceptables por ir en contra de los valores fundamentales, tales como:

- Uso de técnicas subliminales o manipuladoras para alterar el comportamiento de forma que cause perjuicios considerables.
- Explotación de vulnerabilidades de personas (edad, discapacidad, situación socioeconómica) para causar perjuicios significativos.
- Sistemas de "puntuación ciudadana" que evalúen el comportamiento social para generar trato perjudicial injustificado.
- Evaluación de riesgos de criminalidad basada *únicamente* en el perfilado o rasgos de personalidad.
- Creación o ampliación de bases de datos de reconocimiento facial mediante extracción no selectiva de imágenes de internet o CCTV.
- Sistemas de reconocimiento de emociones en el lugar de trabajo y centros educativos (con excepciones médicas o de seguridad).
- Uso de sistemas de identificación biométrica remota "en tiempo real" en espacios públicos por parte de las autoridades, salvo en situaciones excepcionales y bajo estrictas salvaguardias.

8. **Clasificación de Alto Riesgo:** Los sistemas de IA se clasifican como de "alto riesgo" si son componentes de seguridad en ciertos productos (ej. máquinas, dispositivos médicos, automoción, aviación), o si se utilizan en ámbitos críticos como la biometría, infraestructuras críticas, educación y formación profesional, empleo y gestión de trabajadores, acceso a servicios esenciales públicos y privados, garantía del cumplimiento del Derecho, migración y control fronterizo, y administración de justicia y procesos democráticos. Existen excepciones para sistemas de alto riesgo que no influyen sustancialmente en la toma de decisiones o realizan tareas preparatorias/limitadas, a menos que realicen perfilado de personas físicas.

9. **Modelos de IA de Uso General (GPAI):** Los proveedores de GPAI tienen obligaciones específicas de transparencia, incluyendo la elaboración y actualización de documentación técnica y la provisión de información para que otros proveedores puedan integrar el modelo y cumplir con las regulaciones. Deben establecer directrices para cumplir con el derecho de autor, incluyendo la detección y respeto de las reservas de derechos; y hacer público un resumen detallado del contenido utilizado para el entrenamiento del modelo. Los GPAI con **riesgo sistémico** (aquellos con capacidades muy elevadas o gran impacto en el mercado) están sujetos a obligaciones adicionales, como la evaluación continua y la mitigación de riesgos sistémicos (incluyendo pruebas de simulación de adversarios), vigilancia posterior a la comercialización y un nivel adecuado de ciberseguridad. Se fomenta la elaboración de códigos de buenas prácticas para los GPAI.

10. **Vigilancia Posterior a la Comercialización (Post-Market Surveillance):** Los proveedores deben establecer un sistema para recopilar, documentar y analizar activamente la experiencia con el uso de los sistemas de IA de alto riesgo después de su introducción en el mercado, para detectar la necesidad de medidas correctivas y garantizar el cumplimiento continuo. Se deben notificar los incidentes graves a las autoridades competentes.

11. **Apoyo a la Innovación:** Se fomenta la creación de "espacios controlados de pruebas para la IA" (sandboxes regulatorios) a nivel nacional para facilitar el desarrollo y la prueba de sistemas de IA innovadores bajo supervisión regulatoria y durante un tiempo limitado. Se permiten las pruebas de sistemas de IA de alto riesgo en condiciones reales fuera de estos sandboxes, bajo ciertas condiciones y salvaguardias, como el consentimiento informado de los sujetos y la minimización de riesgos. Se busca reducir la carga administrativa y los costes para las PYMES y startups, ofreciendo acceso prioritario a sandboxes y modelos de documentación simplificados.

12. **Gobernanza y Aplicación:** Se establecen estructuras de gobernanza a nivel de la Unión Europea, como la Oficina de IA (AI Office) y el Consejo de IA (AI Board), para coordinar la aplicación y supervisión del Reglamento. Los Estados Miembros deben designar autoridades nacionales competentes (autoridades de notificación y de vigilancia del mercado) con los recursos y conocimientos técnicos necesarios para supervisar la aplicación. Se establecen mecanismos de evaluación de la conformidad, que en el caso de los sistemas de alto riesgo, pueden implicar la participación de organismos notificados externos. Se prevén sanciones efectivas, proporcionadas y disuasorias por el incumplimiento de las obligaciones. Se garantiza el derecho de las personas afectadas a presentar quejas y, en ciertos casos, a obtener una explicación de las decisiones tomadas por la IA.

En resumen, el diseño de herramientas de IA debe ser un proceso multidimensional que no solo considere la innovación y la capacidad técnica, sino que integre de manera fundamental la ética, la seguridad, la transparencia, la equidad y la protección de los derechos humanos desde la concepción hasta la implementación y el monitoreo continuo.

¿CÓMO SE CALCULA EL ÍNDICE DEL RADAR DE LA IA?

Para evaluar el radar de la IA, cada aplicación recibe observaciones cualitativas así como un valor numérico entre [0, 5]. Las puntuaciones numéricas de los ejes del radar se pueden representar como un vector entre estos valores. Normalizando la longitud del vector se obtiene un índice para el radar. Un ejemplo de este procedimiento con la ecuación es:

$$r = \frac{\sqrt{(3.0)^2 + (4.0)^2 + (3.5)^2 + (3.0)^2 + (2.0)^2}}{\sqrt{125}}$$

$$r = \frac{\sqrt{50.25}}{\sqrt{125}} = 0.63$$

El índice del radar r ∈ [0.0, 1.0] considera los cinco ejes para la aplicación seleccionada, el máximo valor es 1. Es posible comparar varias aplicaciones, en el mismo dominio, utilizando los índices.

¿CÓMO SE ANALIZAN LOS EJES DEL RADAR DE LA IA?

A continuación, se presentan algunas preguntas orientadoras para facilitar la evaluación de cada uno de los ejes del Radar de la IA:

1. Eje Diseño Ético:

Las preguntas orientadoras propuestas para diseño ético son:

- ¿El sistema de IA fue diseñado para evitar sesgos en la selección de candidatos, eliminando variables como género, raza o edad que podrían llevar a discriminación?

- ¿Se utilizaron datos equilibrados y representativos en el diseño del sistema de IA?
- ¿El algoritmo es transparente en cómo toma decisiones?
- ¿Cómo se asegura que el sistema priorice la seguridad humana y esté libre de sesgos que puedan afectar negativamente a los trabajadores?
- ¿Se han tenido en cuenta los criterios éticos y la responsabilidad social en el diseño del sistema inteligente?

Así, se propone el siguiente esquema de calificación:

- Calificación 0 - Inexistente/Negligente: El sistema de IA no considera ni aborda los principios éticos en su diseño. Presenta sesgos significativos, falta de transparencia o potencial de daño que no se ha mitigado. Podría estar involucrado en prácticas prohibidas explícitamente por la regulación, como manipulación, explotación o categorización biométrica sensible injustificada. La ciberseguridad y la robustez son totalmente ignoradas, dejando el sistema vulnerable a errores y ataques.

- Calificación 1 - Básico/Reactivo: Se han considerado algunos principios éticos básicos, pero de manera reactiva o superficial, sin un enfoque sistemático de "privacidad desde el diseño". Se realizan esfuerzos mínimos para evitar sesgos obvios o proteger la privacidad, sin una gestión adecuada de la calidad de los datos para prevenir la discriminación. La robustez y la ciberseguridad son deficientes, lo que aumenta el riesgo de fallos y usos indebidos.

- Calificación 2 - Parcial/Consciente: El diseño incorpora activamente algunas consideraciones éticas, como la búsqueda de no discriminación y la privacidad, pero la implementación puede ser inconsistente o incompleta. Se han intentado medidas para garantizar la precisión y la robustez, pero sin alcanzar un nivel adecuado según el estado del arte. Se consideran los valores de la Unión de forma declarativa, pero sin una operacionalización completa.

- Calificación 3 - Adecuado/Conforme: El diseño del sistema de IA incorpora un enfoque sistemático en la ética, alineándose con los principios de la IA fiable (supervisión humana, solidez técnica, gestión de datos, transparencia, diversidad, no discriminación y equidad, bienestar social y ambiental). Los datos de entrenamiento, validación y prueba son de alta calidad, representativos y se han tomado medidas para mitigar sesgos conocidos. Se demuestran niveles adecuados de precisión, robustez y ciberseguridad mediante pruebas apropiadas, abordando vulnerabilidades específicas de la IA. Las prácticas prohibidas están claramente evitadas.

- Calificación 4 - Proactivo/Mejorado: El diseño no solo cumple con los requisitos éticos y de seguridad, sino que va más allá, implementando soluciones innovadoras para anticipar y mitigar nuevos riesgos éticos (p. ej., transparencia activa, explicabilidad avanzada, resistencia a ataques sofisticados). Se demuestra un compromiso continuo con la mejora de la equidad y la explicabilidad, incluso ante escenarios complejos. El sistema facilita activamente la supervisión y la intervención humana.

- Calificación 5 - Ejemplar/Líder: El sistema de IA es un modelo de diseño ético, incorporando las mejores prácticas y contribuyendo a avanzar el estado del arte en IA fiable y centrada en el ser humano. El diseño minimiza proactivamente todos los riesgos conocidos y anticipa los emergentes, siendo altamente adaptable

y resiliente. Sirve de referencia para la industria en la integración de valores éticos y sociales, como la inclusión y la protección de grupos vulnerables, desde la concepción hasta el despliegue.

2. Eje Alfabetización en IA:

Las preguntas orientadoras propuestas para alfabetización son:

- ¿Se aseguran de que los usuarios finales comprendan cómo funcionan los sistemas de IA?
- ¿Qué datos se utilizan y cómo se toman las decisiones?
- ¿Se proporciona capacitación a los usuarios sobre el proceso y las implicaciones de la IA?¿Cuáles son las limitaciones y capacidades del sistema de IA que se explican a los usuarios?
- ¿Cómo se informa al público sobre cómo funciona el sistema, qué datos se recopilan y cómo se utilizan?

Así, se propone el siguiente esquema de calificación:

- Calificación 0 - Nula: No hay conocimiento ni comprensión de los conceptos, beneficios, riesgos, salvaguardias, derechos u obligaciones relacionados con la IA entre el personal o los usuarios. No se llevan a cabo iniciativas de formación o sensibilización.

- Calificación 1 - Incipiente: Se reconoce la necesidad de la alfabetización en IA, pero los esfuerzos son esporádicos o informales. La comprensión es muy básica y no cubre los aspectos técnicos, éticos o regulatorios relevantes para el rol.

- Calificación 2 - Básica/Reactiva: Se han implementado algunas actividades de formación o se han distribuido materiales informativos. El personal relevante tiene una comprensión superficial de los conceptos clave de IA, incluyendo algunos riesgos generales y derechos básicos, pero sin la profundidad necesaria para una toma de decisiones informada o para cumplir con las obligaciones específicas. Se utilizan instrumentos de evaluación para medir los niveles de alfabetización, pero de forma limitada.

- Calificación 3 - Adecuada/Suficiente: Se establecen programas de formación y sensibilización estructurados y se garantiza que el personal (proveedores, responsables del despliegue) tenga un nivel suficiente de alfabetización en IA, adaptado a su función y al contexto de uso del sistema. Se cubren los conceptos necesarios para la toma de decisiones informada, la comprensión de resultados y la identificación de impactos en las personas. La alfabetización se evalúa de manera regular para identificar y abordar las brechas de conocimiento.

- Calificación 4 - Proactiva/Extensa: La organización invierte significativamente en programas de alfabetización en IA que son continuos, completos y adaptados a diversas audiencias (empleados, usuarios, etc.). Se fomenta una cultura de aprendizaje y comprensión profunda de la IA, incluyendo aspectos técnicos, éticos y regulatorios. Se promueve activamente la conciencia pública sobre los beneficios y riesgos de la IA. Se colabora en el desarrollo de marcos de alfabetización.

- Calificación 5 - Ejemplar/Líder: La organización es un referente en la promoción de la alfabetización en IA, desarrollando y compartiendo metodologías y herramientas innovadoras para educar a amplios

segmentos de la sociedad o la industria. El nivel de comprensión de la IA es excepcional en toda la cadena de valor, lo que permite una aplicación óptima de las salvaguardias, el cumplimiento normativo y la identificación proactiva de oportunidades y riesgos emergentes, fomentando la IA fiable.

3. Eje Uso Ético:

Las preguntas orientadoras propuestas para uso ético son:

- ¿Existen protocolos para que los usuarios revisen las decisiones de la IA y corrijan posibles errores?
- ¿Se limita el uso de la IA a fines legítimos y con supervisión humana?
- ¿Cómo se asegura que la IA complemente el trabajo humano en lugar de reemplazarlo abruptamente?
- ¿Se han establecido protocolos estrictos para el uso del sistema y se garantiza la supervisión humana?
- ¿Se evita el uso de la IA para vigilancia masiva o invasión de la privacidad?

Así, se propone el siguiente esquema de calificación:

- Calificación 0 - Prohibido/No conforme: El uso del sistema de IA cae directamente en las prácticas prohibidas por la regulación (p. ej., manipulación subliminal, puntuación ciudadana discriminatoria, identificación biométrica remota en espacios públicos sin justificación legal). No se respetan los derechos fundamentales de los individuos o se ignoran las obligaciones regulatorias básicas.

- Calificación 1 - Problemático/Mínimo: El uso evita las prohibiciones explícitas más flagrantes, pero se realiza con una comprensión mínima de las implicaciones éticas y de derechos fundamentales. La supervisión humana es deficiente o inexistente, y no se informa adecuadamente a los usuarios o personas afectadas sobre el uso de la IA.

- Calificación 2 - Reactivo/Inconsistente: Se reacciona a los problemas éticos a medida que surgen, en lugar de prevenirlos. El cumplimiento de las obligaciones de uso ético es inconsistente. La información a las personas afectadas es básica, y la supervisión humana se realiza de manera ad-hoc o sin la formación adecuada. No se abordan los riesgos de "sesgo de automatización".

- Calificación 3 - Adecuado/Conforme: El uso del sistema de IA cumple con todas las obligaciones regulatorias relativas a las prácticas permitidas y prohibidas. Se garantiza una supervisión humana efectiva y proporcionada al riesgo, con personal competente y capacitado para entender, interpretar e intervenir en el sistema. Se informa a las personas afectadas sobre el uso de la IA y sus derechos, como el derecho a la explicación. En caso de generación de contenido, se cumplen las obligaciones de marcado/divulgación.

- Calificación 4 - Proactivo/Responsable: Se adoptan medidas proactivas para asegurar que el uso del sistema de IA sea ético y beneficie a la sociedad, más allá del mero cumplimiento. Se implementan salvaguardias adicionales para proteger a grupos vulnerables y se busca activamente la no discriminación. La supervisión humana es robusta, con mecanismos claros de intervención y reversión. Se fomenta la transparencia y la explicabilidad en el uso.

- Calificación 5 - Ejemplar/Innovador: El uso del sistema de IA no solo es ético y conforme, sino que establece un estándar en la industria o sector, demostrando un compromiso inquebrantable con los valores humanos y los derechos fundamentales. Se desarrollan e implementan continuamente mejores prácticas para garantizar que el uso de la IA contribuya al bienestar social y ambiental, minimizando cualquier efecto adverso y adaptándose a nuevos desafíos éticos. Se realizan evaluaciones de impacto sobre derechos fundamentales de forma rigurosa y se aprende de ellas.

4. Eje Manejo de Riesgos:

Las preguntas orientadoras propuestas para manejo de riesgos son:

- ¿Qué riesgos se han identificado, como la posibilidad de que el sistema refuerce sesgos existentes o tome decisiones incorrectas?
- ¿Se implementan pruebas continuas para detectar sesgos y errores?
- ¿Existe un plan de contingencia para corregir los riesgos identificados?
- ¿Cómo se evalúa y administra los riesgos, considerando su impacto, plausibilidad y el contexto de aplicación de la IA?
- ¿Se han implementado medidas para mitigar riesgos y proteger los datos?

Así, se propone el siguiente esquema de calificación:

- Calificación 0 - Inexistente/Ignorado: No existe un sistema de gestión de riesgos para la IA. Los riesgos no se identifican, evalúan ni mitigan de forma alguna, o se ignoran deliberadamente. No hay monitoreo post-comercialización ni planes de contingencia.

- Calificación 1 - Básico/Fragmentado: Se reconocen algunos riesgos obvios, pero el enfoque es ad-hoc y no sistemático. Las medidas de mitigación son mínimas y no se integran en un marco coherente. La documentación de riesgos es escasa o inexistente. No hay un monitoreo continuo.

- Calificación 2 - Parcial/Reactivo: Existe un sistema de gestión de riesgos, pero es reactivo y no cubre todos los riesgos relevantes (p. ej., solo técnicos, no éticos/sociales). La identificación de riesgos es incompleta, y la evaluación carece de profundidad. Las medidas de mitigación son implementadas solo después de que los problemas han surgido. El monitoreo post-comercialización es superficial.

- Calificación 3 - Adecuado/Conforme: Se ha establecido, implementado y documentado un sistema de gestión de riesgos iterativo y continuo, cubriendo la identificación, análisis, estimación, evaluación y mitigación de riesgos conocidos y razonablemente previsibles (incluido el uso indebido previsible). Se tienen en cuenta la salud, la seguridad y los derechos fundamentales. Se realiza una revisión y actualización periódica del sistema. Se cuenta con un plan de vigilancia post-comercialización, y se realizan informes de incidentes graves.

- Calificación 4 - Proactivo/Integrado: El sistema de gestión de riesgos es robusto, proactivo e integrado en todas las fases del ciclo de vida de la IA. Se anticipan y abordan los riesgos emergentes, incluyendo aquellos de "modelos" de IA y sus interacciones complejas. Se realizan pruebas rigurosas para verificar las medidas

de mitigación. Se considera explícitamente el impacto en grupos vulnerables y menores. La documentación es exhaustiva y se involucra a expertos externos. Se implementan ciberseguridad avanzada y resiliencia contra ataques específicos de IA.

- Calificación 5 - Ejemplar/Holístico: La gestión de riesgos de IA es un modelo de excelencia, aplicando un enfoque holístico que abarca no solo los requisitos técnicos y normativos, sino también las implicaciones éticas y sociales a largo plazo. Se implementan sistemas de detección temprana de riesgos y se contribuye a la estandarización y mejora de las metodologías de gestión de riesgos a nivel de la industria (p. ej., el framework KAIRI SAFE). Se promueve una cultura de responsabilidad y seguridad en toda la organización y se comparte el conocimiento con la comunidad.

5. Eje Regulación:

Las preguntas orientadoras propuestas para regulación son:

- ¿El sistema cumple con las normativas locales e internacionales sobre privacidad y no discriminación?
- ¿Se asegura de que el sistema esté auditado y certificado para cumplir con todas las regulaciones aplicables?¿El sistema cumple con todas las leyes de privacidad y protección de datos?
- ¿Cómo se asegura de que el sistema cumpla con las normativas laborales y de seguridad industrial a nivel nacional e internacional?
- ¿Se han considerado las diferentes aproximaciones regulatorias, como el Reglamento de IA de la Unión Europea o las directrices de la OCDE?

Así, se propone el siguiente esquema de calificación:

- Calificación 0 - Ausente/Ignorada: No se ha tomado ninguna medida para cumplir con las regulaciones de IA existentes o emergentes, o se actúa en abierta contravención de ellas.

- Calificación 1 - Mínima/Conocimiento básico: Se tiene un conocimiento muy limitado de las regulaciones aplicables. Los esfuerzos de cumplimiento son escasos, se centran solo en los aspectos más básicos y no se adaptan a la clasificación de riesgo de los sistemas de IA.

- Calificación 2 - Parcial/Interpretación limitada: Se reconocen las regulaciones, pero la interpretación es limitada o sesgada. Se intenta cumplir, pero con brechas significativas, especialmente para sistemas de alto riesgo o modelos de propósito general. La documentación es insuficiente, y no se participa en los mecanismos de gobernanza o certificación.

- Calificación 3 - Adecuada/Plena Conformidad: Se cumple con todos los requisitos aplicables de la regulación de IA (AI Act), incluyendo la clasificación de sistemas de alto riesgo, las prohibiciones específicas, los requisitos obligatorios (gestión de riesgos, datos, documentación, transparencia, supervisión humana, robustez, ciberseguridad) y las obligaciones de los proveedores y responsables del despliegue. Se utilizan estándares armonizados o especificaciones comunes cuando están disponibles. Los sistemas se registran en la base de datos de la UE cuando es necesario. Se aplican sanciones proporcionadas y se responde a los incidentes.

- Calificación 4 - Proactiva/Participativa: La organización no solo cumple con la regulación, sino que participa activamente en su desarrollo y mejora. Se interactúa con las autoridades reguladoras (Consejo de IA, Oficina de IA), se contribuye a la elaboración de códigos de conducta y buenas prácticas, y se prepara proactivamente para futuras actualizaciones normativas. Se implementan medidas que van más allá del mínimo legal, anticipando las expectativas regulatorias.

- Calificación 5 - Ejemplar/Líder Global: La organización es un referente en la implementación regulatoria de la IA, influyendo positivamente en el desarrollo de políticas y estándares a nivel global. Se contribuye significativamente a la seguridad jurídica y a la adopción de una IA fiable en la Unión, por ejemplo, mediante la participación en sandboxes regulatorios y pruebas en condiciones reales de forma ejemplar. Se promueve activamente la coherencia y la armonización de las prácticas regulatorias en la UE y a nivel internacional.

EJEMPLOS DE APLICACIÓN DEL RADAR DE LA IA

A continuación, se presentan algunos ejemplos de cómo aplicar el radar de la IA en diferentes contextos:

1. **Ejemplo 1:** Sistema de reclutamiento basado en IA. En este caso los ejes se analizarían como:

   - Diseño Ético: Evaluar si el sistema está diseñado para evitar sesgos en la selección de candidatos, eliminando variables como género, raza o edad que podrían llevar a discriminación. Utilizar datos equilibrados y representativos, y asegurar que el algoritmo sea transparente en cómo toma las decisiones.

   - Alfabetización en IA: Asegurarse de que los reclutadores y los candidatos entiendan cómo funciona el sistema, qué datos se utilizan y cómo se toman las decisiones. Proporcionar capacitación a los reclutadores y transparencia a los candidatos sobre el proceso de selección.

   - Uso Ético: Evitar el uso de la IA para descartar candidatos sin una revisión humana adicional. Establecer protocolos para que los reclutadores revisen las decisiones de la IA y corrijan posibles errores.

   - Manejo de Riesgos: Identificar los riesgos, como la posibilidad de que el sistema refuerce sesgos existentes o tome decisiones incorrectas. Implementar pruebas continuas para detectar sesgos y errores, y tener un plan de contingencia para corregirlos.

   - Regulación: Verificar si el sistema cumple con las normativas locales e internacionales sobre privacidad y no discriminación. Asegurarse de que el sistema esté auditado y certificado para cumplir con todas las regulaciones aplicables.

2. **Ejemplo 2**: Sistema de reconocimiento facial utilizado en espacios públicos. En este caso los ejes se analizarían como:

- Diseño Ético: Asegurarse de que el sistema esté diseñado para minimizar falsos positivos y evitar la discriminación por raza, género u otras características. Utilizar algoritmos entrenados con datos diversos y equilibrados, y asegurar que el sistema tenga una alta precisión en la identificación.

- Alfabetización en IA: Informar a las autoridades y al público sobre cómo funciona el sistema, qué datos se recopilan y cómo se utilizan. Proporcionar información clara y accesible sobre el funcionamiento del sistema, y educar a los ciudadanos sobre sus implicaciones.

- Uso Ético: Utilizar el sistema solo para fines legítimos, como la seguridad pública, y no para vigilancia masiva o invasión de la privacidad. Establecer protocolos estrictos para el uso del sistema y garantizar la supervisión humana.

- Manejo de Riesgos: Implementar medidas para mitigar riesgos y proteger datos.

- Regulación: Cumplir con todas las leyes de privacidad y protección de datos.

Ejemplo 3: Aplicación de IA para la supervisión automatizada de líneas de producción en una fábrica.

- Diseño Ético: Asegurarse de que el sistema priorice la seguridad humana y esté libre de sesgos que puedan afectar negativamente a los trabajadores.

- Alfabetización en IA: Capacitar a los trabajadores en el uso del sistema y explicarles sus limitaciones y capacidades.

- Uso Ético: Implementar políticas que aseguren que la IA complemente el trabajo humano en lugar de reemplazarlo abruptamente.

- Manejo de Riesgos: Establecer protocolos de seguridad y realizar pruebas continuas para garantizar la fiabilidad del sistema.

- Regulación: Asegurarse de que el sistema cumpla con las normativas laborales y de seguridad industrial en Colombia y a nivel internacional.

Ejemplo 4: Diseño de un arma automatizada con IA.

- Diseño Ético: Diseñado para realizar ataques contra seres humanos, incluyendo la capacidad de decidir autónomamente con los objetivos establecidos.

- Alfabetización en IA: No brinda información al usuario de sus modelos de IA, únicamente sus características técnicas y capacidades. Brinda poca información sobre el uso de la IA en el producto y su funcionamiento no requiere ningún tipo de capacitación.

- Uso Ético: Tener la capacidad de detectar rostros, autonomía y puede incluir armas de gran alcance, por lo que su uso puede incurrir en graves faltas éticas.

- Manejo de Riesgos: Aunque asegura que siempre debe haber un humano en el loop de funcionamiento, se han registrado ataques autónomos sin claridad de sus consecuencias.

- Regulación: No cuenta con una regulación clara. El dispositivo cumple regulaciones técnicas asociados a su efecto en la salud humana, pero no sigue un marco regulatorio claro sobre su uso en aplicaciones militares.

Ejemplo 5: Creación de una herramienta de educación basada en IA. En este caso los ejes se analizarían como:

- Diseño Ético: La plataforma está diseñada para mejorar el aprendizaje de los estudiantes de manera individualizada, fortaleciendo sus capacidades y debilidades.

- Alfabetización en IA: Debe haber suficiente documentación de los sistemas tanto para los estudiantes como para los profesores. Asegura transparencia de los modelos de IA usada para los educadores que la usan, además de brindar información completa para los estudiantes.

- Uso Ético: Recepción de múltiples reconocimientos por sus beneficios para el ciclo de aprendizaje de los estudiantes de educación escolar y superior.

- Manejo de Riesgos: No presentar mayores riesgos, asegura la privacidad de los usuarios y tiene múltiples herramientas para informar sobre su correcta utilización. Tiene estrictas herramientas de privacidad y seguridad, asegura una revisión constante para evitar sesgos y da control al educador sobre las recomendaciones.

- Regulación: Seguir las regulaciones y acuerdos relativos a la educación a nivel mundial, teniendo alianzas con universidades y gobiernos que la destacan en muchos países. Sigue las normativa y regulación de educación motivada por diferentes acuerdos internacionales, siendo la herramienta oficial de educación de gobiernos e importantes centros.

Ejemplo 6: Aplicación de IA para la detección de noticias falsas. En este caso los ejes se analizarían como:

- Diseño Ético: Seguir estándares como precisión, imparcialidad y transparencia, combinando IA con supervisión humana para verificar información y corrección de errores. Diseñada para informar al público sobre posibles noticias falsas en la web de manera imparcial y en favor de un buen uso de la IA en medios de comunicación.

- Alfabetización en IA: Capacitar a los periodistas en el uso del sistema y explicarles sus limitaciones y capacidades. Brinda información completa y transparente al usuario de su funcionamiento y el uso de IA, teniendo código abierto, APIs y explicación detallada de algoritmos.

- Uso Ético: Su misión social es combatir la desinformación. Declara cumplir códigos de buenas prácticas y enfatiza que no censura, sino que aclara la veracidad de afirmaciones. Asegura imparcialidad, justicia y

ética en su monitoreo de noticias, incluyendo la intervención humana en todo momento del proceso, pero puede incluir sesgos.

- Manejo de Riesgos: Utilizar la IA en conjunto con analistas humanos para identificar noticias falsas en tiempo real. Cuenta con un protocolo público de corrección de errores. Clasifica los sitios de noticias según los nueve criterios del periodismo responsable, identificando también si la información se presenta como sátira o es oficial.

- Regulación: Se guía por principios éticos y de libertad de expresión, declara adherirse al código de principios de la International Fact-Checking Network (IFCN).

Ejemplo 7: Sistema de IA para apoyar el trabajo policial. En este caso los ejes se analizarían como:

- Diseño Ético diseñada bajo los estándares de seguridad más importantes del mundo para combatir toda amenaza criminal. Ha sido criticado por el uso masivo de imágenes sin consentimiento, falta de transparencia y no tener un enfoque centrado en las personas.

- Alfabetización en IA: Ni los ciudadanos ni las agencias que la usan comprenden completamente cómo funciona, ya que no hay esfuerzos significativos de la empresa por educar a los usuarios en torno a la plataforma. Solo brinda información profunda de algoritmos a las organizaciones con cursos.

- Uso Ético: Tiene acceso a redes sociales y bases de datos de la policía, por lo que puede violar la privacidad de datos. Su aplicación es polémica y ha sido asociada a vigilancia masiva sin supervisión, violando los derechos fundamentales a pesar de ser utilizada por la policía.

- Manejo de Riesgos: No tiene planteamientos de manejo de riesgos para la sociedad, notando que dan poca relevancia a las consecuencias legales y éticas de la aplicación y su uso.

- Regulación: Cumple todas las normativas vigentes de EE.UU. como la CCPA, COPPA, el acta de protección de datos y de seguridad de la UE. En muchos lugares donde opera, no cumple con normativas de protección de datos, como el GDPR en Europa, y ha sido prohibido en varias jurisdicciones.

Ejemplo 8: Uso de IA para el espionaje cibernético. En este caso los ejes se analizarían como:

- Diseño Ético: Busca aprovechar las vulnerabilidades desconocidas de diferentes sistemas operativos para realizar vigilancia. No incorpora principios de justicia, transparencia ni respeto por los derechos humanos en su diseño, su objetivo es la vigilancia.

- Alfabetización en IA: Funciona de manera secreta y el desarrollador nunca hace mención directa a ella, por lo que no se sabe cómo funciona y es difícil de detectar. No hay esfuerzo alguno para educar o informar sobre su funcionamiento ni sus implicaciones, actúa de manera secreta.

- Uso Ético: Está pensada para atacar cualquier dispositivo móvil, por lo que ha sido empleada para periodistas, activistas y figuras políticas. Está pensada para atacar cualquier dispositivo con sistema Microsoft o Google, por lo que ha sido usada para espiar activistas y disidentes.

- Manejo de Riesgos: No tiene ningún tipo de manejo de riesgos, aprovechando medios de uso habitual para penetrar en los dispositivos y robar información. Fue implementado sin control de daños ni medidas para mitigar impactos negativos, por lo que ha sido incluido en varias listas negras.

- Regulación: Opera al margen de marcos regulatorios efectivos y viola las leyes de privacidad de los países donde ha sido encontrado. No sigue ningún marco ético, los desarrolladores han recibido múltiples demandas por su accionar y es ilegal en la mayoría de los países.

Ejemplo 9: Herramienta de IA para la selección de empleo. En este caso los ejes se analizarían como:

- Diseño Ético: Asegura la eliminación de sesgos con su enfoque en habilidades blandas, pero puede afectar indirectamente a los candidatos. Integra valores como transparencia y reducción de sesgos con el genoma profesional, pero sin una explicación abierta del funcionamiento interno.

- Alfabetización en IA: Ofrece algunos elementos explicativos sobre su sistema, pero no educa activamente al usuario sobre IA o toma de decisiones.

- Uso Ético: La herramienta puede motivar discriminación por parte de los reclutadores o una limitación en cómo se presentan los candidatos. Promueve usos legítimos y responsables, centrados en el talento humano. No se evidencian prácticas discriminatorias.

- Manejo de Riesgos: Evalúa múltiples factores para mejorar el matching, pero no tiene mecanismos de revisión pública ni reportes de errores al usuario. Da el control de la privacidad al usuario y aseguran que el control automatizado evita sesgos, pero se basa en un sistema de puntuación.

- Regulación: Se rige bajo la legislación de Ontario, Canadá, para aplicaciones de búsqueda de empleo y establecimiento de ofertas reales. No comunica estrategias claras frente a marcos regulatorios de IA.

Ejemplo 10: Asistente personal basado en IA. En este caso los ejes se analizarían como:

- Diseño Ético: Es transparente para el usuario, pero únicamente aseguran no tomar datos de sus clientes. Busca minimizar la recolección de datos personales, integrando principios éticos desde su diseño.

- Alfabetización en IA: Documentación accesible que explica cómo se entrenan y funcionan sus modelos, dando al usuario una comprensión general de su funcionamiento y límites. Cuenta con una buena cantidad de información sobre el entrenamiento y funcionamiento de los modelos para el usuario.

- Uso Ético: Aunque afirma proteger la privacidad de la información, han habido cuestionamientos sobre el uso y almacenamiento de datos.

- Manejo de Riesgos: Cuenta con protocolos robustos de seguridad, pero no está exento de incidentes. Se han reportado vulnerabilidades y posibles usos indebidos.

- Regulación: Cumplen con la regulación de EE.UU. y la Ley de la IA de la UE, además de tener enfoque en la IA responsable. Muestra compromiso con prácticas de IA responsable, pero su enfoque está condicionado por intereses corporativos.

¿EN QUÉ SE BASA SCIENTIA? ELIZA Y SU RELEVANCIA

El desarrollo tiene su origen en ELIZA, que es un programa informático influyente desarrollado en la década de 1960 por Joseph Weizenbaum, profesor de Ciencias de la Computación en el Instituto Tecnológico de Massachusetts (MIT). Es considerado el precursor de todas las interfaces conversacionales y chatbots.

- Creación y Propósito: Weizenbaum desarrolló ELIZA para simular una conversación, inicialmente concibiéndola como un barman antes de decidir que los psiquiatras serían más interesantes. Su personalidad más famosa fue DOCTOR, un *script* diseñado para parodiar a un psicoterapeuta rogeriano. El enfoque rogeriano se basa en una aproximación lingüística no directiva, donde el cliente hace la mayor parte de la conversación. ELIZA respondía a las entradas de texto en inglés con respuestas escritas, a menudo reformulando elementos clave de la entrada del usuario en forma de preguntas.

- Funcionamiento y Mecanismo: Utilizaba técnicas de coincidencia de patrones (pattern-matching) relativamente simples para transformar las oraciones de entrada en respuestas que parecían tener sentido para el usuario. Escaneaba la oración de entrada de izquierda a derecha, buscando palabras clave en un diccionario. Si se identificaba una palabra clave, solo se intentaban las reglas de descomposición que contenían esa palabra clave. Además, podía dar un tratamiento único a ciertas palabras clave; por ejemplo, si alguien usaba un universal como "todos" o "siempre", ELIZA podría responder con "De quién en particular estás pensando". Weizenbaum diseñó el sistema para que el conocimiento de un dominio (como la alta costura o el béisbol) residiera en un módulo de programa separado de la parte que manejaba las conversaciones. Esto permitía que el programa hablara sobre una variedad de temas simplemente cargando el módulo de *software* adecuado.

- Origen del Nombre: ELIZA fue nombrada así por Eliza Doolittle, un personaje de la obra *Pygmalion* de George Bernard Shaw. Weizenbaum se inspiró en la adaptación musical *My Fair Lady*. El nombre también aludía a la idea de que, al igual que Miss Doolittle, nunca estaba claro si el programa se volvía más inteligente.

- Tecnología de Programación: Weizenbaum comenzó su desarrollo de ELIZA mientras desarrollaba un sistema de programación llamado Symmetric List Processor (SLIP). La versión original de ELIZA fue creada en MAD-SLIP (una combinación de MAD, Michigan Algorithm Decoder, y SLIP) en una computadora IBM 7094 en el MIT. Existe una idea errónea persistente de que ELIZA fue escrita originalmente en LISP, en parte debido a la circulación de una conversión de LISP por Bernie Cosell.

- Versiones y Documentación: Weizenbaum continuó trabajando en ELIZA, y las versiones posteriores incorporaron funciones más sofisticadas, como un evaluador de expresiones de complejidad ilimitada y la

capacidad de manejar hasta tres *scripts* simultáneamente. Se han identificado al menos cinco versiones principales de ELIZA, aunque gran parte del código fuente original de las versiones posteriores se ha perdido o está pobremente documentado.

- Impacto y "Efecto ELIZA": Weizenbaum se sorprendió por la rapidez y profundidad con la que las personas que conversaban con DOCTOR se involucraban emocionalmente con la computadora y la antropomorfizaban. Esto llevó a la identificación del "efecto ELIZA", la tendencia de los humanos a atribuir comprensión e inteligencia a los sistemas informáticos, incluso cuando no están garantizadas. A pesar de su aparente simplicidad hoy en día, ELIZA fue un hito en la historia de la informática y planteó cuestiones filosóficas, sociales y políticas cruciales sobre la interacción entre humanos y máquinas.

Weizenbaum pasó de ser un entusiasta investigador de la IA a uno de sus críticos más reconocidos debido a su preocupación por cómo se percibía ELIZA y sus implicaciones. Le impactó que los psicoterapeutas pensaran que podía automatizar la terapia, que los usuarios la trataran como un ser humano, y la creencia de que podía generalizarse para entender el lenguaje natural en su totalidad. ELIZA también tuvo un impacto en la cultura popular, apareciendo, por ejemplo, en el álbum de 1971 "I Think We're All Bozos on This Bus" de Firesign Theatre, y siendo referenciada por el chatbot Siri de Apple.

ELIZA planteó muchos de los riesgos y preocupaciones que hoy son centrales en el debate sobre la inteligencia artificial (IA) moderna, como ChatGPT. Los riesgos asociados a los chatbots, tanto los tempranos como los actuales, se pueden clasificar en varias categorías clave:

1. **Atribución de comprensión e inteligencia no justificada (Efecto ELIZA):** Weizenbaum se sorprendió por la rapidez y profundidad con la que las personas que interactuaban con ELIZA (especialmente con su *script* DOCTOR, que simulaba un psicoterapeuta rogeriano) se involucraban emocionalmente con la computadora y la antropomorfizaban.

   Este fenómeno se conoció como el "efecto ELIZA": la tendencia de los humanos a atribuir comprensión e inteligencia a los sistemas informáticos, incluso cuando no está justificado. Hofstadter lo describió como la "susceptibilidad de las personas a interpretar mucha más comprensión de la justificada en las cadenas de símbolos -especialmente palabras- unidas por ordenadores". Los usuarios llegaban a revelar secretos y atribuir cualidades humanas al sistema técnico.

   Weizenbaum se preocupó de que la realidad interna de las personas pudiera ser reemplazada por la de la máquina. Esta "pretensión de simpatía o respeto interpersonal" por parte del ordenador, al seguir solo una lógica programada, podía transformar un encuentro comunicativo potencialmente transformador en uno alienante. Incluso con sistemas actuales como ChatGPT, que utilizan frases en primera persona y simulan formas de cortesía humana, el usuario debe recordar consistentemente que las respuestas son generadas artificialmente.

2. **Falta de "verdadera comprensión" y generación de información errónea:** ChatGPT mismo admite que "no posee conocimiento de la misma manera que los humanos" y "no entiende realmente los conceptos de la forma en que lo hacen los humanos", sino que "reconoce e imita patrones de lenguaje, información y

contexto" de sus datos de entrenamiento. Su comprensión se basa en la semántica inferencial, no en la semántica referencial (experiencia en el mundo real).

El problema central de ChatGPT es cómo maneja la verdad. En lugar de buscar afirmaciones verdaderas, el sistema se preocupa por la probabilidad de que una afirmación sea "precisa", lo que es un sustituto probabilístico de la verdad basado en la correlación textual. Esto lleva a que los chatbots modernos puedan producir una mezcla de resultados claros y precisos y desinformación arbitraria sin clarificar su propio alcance. Ejemplos incluyen dar información biográfica incorrecta y contradictoria sobre personas (como el caso de Christiane Floyd).

El conocimiento generado por ChatGPT se describe como "conocimiento molido": no tiene una fuente clara ni autoría, y es volátil, lo que significa que el sistema puede dar una respuesta diferente a la misma pregunta en otro momento. Esto impide usarlo en una argumentación seria. Los usuarios no pueden confiar en los resultados sin verificarlos, a menudo recurriendo a motores de búsqueda o enciclopedias en línea para contrastar la información. Existe el concepto de "estupidez artificial", donde los algoritmos de aprendizaje automático cometen errores "estúpidos" o son simplificados para parecer más humanos. Estos errores no son como los errores humanos; no son eventos de aprendizaje para el sistema y no conducen a una mejora genuina en su "comprensión".

3. **Sesgo y opacidad:** Las respuestas de ChatGPT pueden exhibir **sesgo cultural y ontológico**, reflejando la perspectiva predominante de los datos de entrenamiento (por ejemplo, una visión angloamericana de la Primera Guerra Mundial).
   A diferencia de los textos de origen humano, donde el sesgo puede atribuirse a autores, los modelos de lenguaje de IA presentan sus respuestas como "generales y objetivas", ocultando un sesgo "implícito y omnipresente".

   Los sistemas de IA modernos son "enormemente complejos" y su funcionamiento interno es difícil de entender, controlar y asegurar que no causen daño. Esto contrasta con ELIZA, que era auto-contenida y basada en reglas, lo que permitía a los especialistas explicar su funcionamiento.

4. **Implicaciones para la responsabilidad y la sociedad:** Weizenbaum se preocupó de que, con la delegación de tareas a las máquinas, "ningún ser humano es ya responsable de 'lo que dice la máquina'". La creciente dependencia de la sociedad en sistemas informáticos que superan la comprensión de sus usuarios es un "desarrollo muy serio". La IA puede "deshumanizar" a los individuos al tratarlos como menos que personas completas, eludiendo contextos humanos que dan significado al lenguaje.

   Las tecnologías de IA, especialmente cuando son impulsadas por el capitalismo, pueden conducir a la automatización de la producción cultural y la transformación del trabajo humano en una mercancía, ocultando la labor social detrás de algoritmos y creando formas de trabajo alienadas. La "obfuscación de los medios" (ya sea mano de obra humana o IA) y la racionalización de los seres humanos en sistemas computacionales pueden crear una relación de "mando-ejecución" entre el usuario y el proceso subyacente.

   El desarrollo de la IA se está subsumiendo cada vez más a las necesidades del capitalismo, con suposiciones implícitas de la superioridad de los mercados para estructurar las relaciones sociales. Los modelos de

lenguaje grandes (LLMs) como ChatGPT requieren "enormes cantidades de poder de procesamiento y tiempo de GPU" para su funcionamiento.

5. **Desafíos para la interacción y el aprendizaje humano:** La dificultad de argumentar con ChatGPT es "profundamente frustrante"; las objeciones a menudo resultan en respuestas "pseudocorteses" o reformulaciones de la propia entrada del usuario, lo que se describe como una "danza degenerada". La falta de transparencia en el proceso de aprendizaje del sistema, la dependencia de la selección del corpus de texto y la imposibilidad de predecir o influir en su base de conocimiento son desafíos significativos. Las limitaciones actuales de ChatGPT exigen que los desarrolladores adopten políticas para reducir la "estupidez artificial".

Existe el riesgo de que los usuarios se contenten con textos generados por IA sin una "cultura de aprendizaje más refinada" que implique leer, discutir, criticar, verificar y revisar el contenido generado. En resumen, los riesgos de los chatbots, desde ELIZA hasta las IA generativas modernas como ChatGPT, radican en su capacidad para generar desinformación, ocultar sesgos, carecer de una verdadera comprensión (a pesar de parecer muy humanos), deshumanizar la interacción, y crear nuevas formas de alienación laboral y social. Weizenbaum ya advertía que no se debe dar a las computadoras tareas que impliquen juicio humano y que es crucial que los profesionales de la informática se pregunten sobre el uso final de su trabajo.

¿QUIÉN FUE JOSEPH WEIZENBAUM?

Joseph Weizenbaum fue un profesor de Ciencias de la Computación en el Instituto Tecnológico de Massachusetts (MIT). Es ampliamente reconocido por desarrollar ELIZA, un influyente programa informático en la década de 1960 que se considera el precursor de todas las interfaces conversacionales y chatbots.
A continuación, se detalla quién era Joseph Weizenbaum y su impacto:

- **Orígenes y Formación:** Joseph Weizenbaum nació en el seno de una familia judía en Berlín el 8 de enero de 1923. A la edad de 13 años, huyó con sus padres de la Alemania nazi y emigró a los Estados Unidos, estableciéndose en Detroit. Sus estudios de matemáticas en la Universidad de Wayne fueron interrumpidos por la guerra, durante la cual sirvió en el servicio meteorológico de la Fuerza Aérea. Regresó a la universidad en 1946, obteniendo su licenciatura en Matemáticas en 1948 y una maestría en 1950.

- **Carrera Temprana y Contribuciones Técnicas:** Ayudó a diseñar y construir una computadora digital en la Universidad de Wayne en Detroit. En 1955, se unió a General Electric en la Costa Oeste para colaborar en el desarrollo de ERMA (Electronic Recording Machine, Accounting), un sistema de automatización bancaria para el Bank of America. Este sistema permitía el procesamiento automatizado de cheques usando tinta magnética y reconocimiento de caracteres, y su demostración inicial incluyó un programador oculto para simular un funcionamiento perfecto, una técnica conocida como "Wizard of Oz". Esta temprana experiencia con el engaño en la computación y el reemplazo del trabajo humano prefiguró sus futuras preocupaciones sobre la automatización. Mientras estuvo en General Electric, Weizenbaum desarrolló el sistema de programación Symmetric List Processor (SLIP), completándolo en 1963. En 1962, escribió un artículo titulado "How To Make a Computer Appear Intelligent", donde argumentaba que una actividad que "produce resultados de una manera que no parece comprensible para un observador particular, le parecerá a ese observador ser de alguna manera inteligente, o al menos motivada inteligentemente". Este trabajo,

que buscaba crear la "ilusión poderosa" de que la computadora era inteligente, fue un precursor de ELIZA y le valió la reputación de "charlatán o estafador".

- **Creación de ELIZA:** En 1963, le ofrecieron un puesto de profesor asociado en ingeniería eléctrica en el MIT debido a su trabajo en SLIP. Allí, reescribió SLIP en MAD (Michigan Algorithm Decoder), contrarrestando la "idea errónea persistente" de que ELIZA fue escrita originalmente en LISP. En el MIT, a mediados de la década de 1960, Weizenbaum buscaba que las computadoras pudieran "hablar con la gente en inglés". Creó ELIZA, un programa diseñado para simular una conversación, que es un "ancestros de todas las interfaces conversacionales y chatbots". La personalidad más famosa de ELIZA fue DOCTOR, un *script* que parodiaba a un psicoterapeuta rogeriano, el cual respondía a las entradas del usuario con preguntas que a menudo reformulaban elementos clave de lo que el usuario había dicho. El nombre ELIZA fue inspirado por Eliza Doolittle de la obra *Pygmalion* y el musical *My Fair Lady*, aludiendo a la idea de que "nunca estuvo del todo claro si el programa se volvía más inteligente".

- **Crítica a la IA:** Weizenbaum se sorprendió por la rapidez y profundidad con la que las personas se involucraron emocionalmente con DOCTOR y lo antropomorfizaron, tratándolo como una persona a la que podían confiar secretos. Esto llevó a la identificación del "efecto ELIZA": la tendencia de los humanos a atribuir comprensión e inteligencia a los sistemas informáticos, incluso cuando no está justificado. La experiencia con ELIZA transformó a Weizenbaum de un investigador entusiasta de la IA a uno de sus críticos más reconocidos. Le preocupaba que los psicoterapeutas creyeran que DOCTOR podría automatizar la terapia, que los usuarios lo trataran como un ser humano, y la creencia de que podía generalizarse para comprender el lenguaje natural en su totalidad. Se opuso a la idea de que los humanos son como máquinas y criticó la "racionalización" y la "calculabilidad de la realidad" que la computación introduce en la sociedad.

  Argumentó que las computadoras no deben recibir tareas que requieran juicio humano porque operan sobre bases que ningún ser humano debería aceptar. Le preocupaba que, al delegar tareas a las máquinas, "ningún ser humano es ya responsable de 'lo que dice la máquina'". Advirtió que la creciente dependencia de la sociedad en sistemas informáticos que superan la comprensión de sus usuarios es un "desarrollo muy serio". Abogó por una "mirada profunda" a la amenaza que los profesionales de la computación tienen el poder de alterar el estado del mundo "de una manera conducente a la vida". En su obra, Weizenbaum criticó la participación de los profesionales de la computación en el desarrollo de sistemas de armas que "amenazan el asesinato a escala genocida". Sostuvo que sin la cooperación de los profesionales de la informática, la carrera armamentística no podría avanzar.

  Denunció el "embellecimiento del lenguaje" (euphemistic linguistic dissimulation) en la investigación de la IA, donde se habla de sistemas que "entienden, ven, deciden, juzgan" sin reconocer la "superficialidad e inmensurable ingenuidad" propia de quienes los crean. Para él, esto "anestesia nuestra capacidad de evaluar la calidad de nuestro trabajo y, lo que es más importante, de identificar y tomar conciencia de su uso final".

- **Legado y Relevancia Actual:** Weizenbaum destacó que el desarrollo de la IA se está subsumiendo cada vez más a las necesidades del capitalismo, con suposiciones implícitas sobre la superioridad de los mercados para estructurar las relaciones sociales. Advirtió que el peligro de los sistemas de IA es que la destrucción de la capacidad del pensamiento humano pueda ser malinterpretada como un "efecto secundario

accidental". Sus preocupaciones sobre la manipulación y el engaño por parte de los sistemas informáticos, y el riesgo de que la empatía humana se desvíe hacia las máquinas con fines de lucro o para socavar la democracia, siguen siendo relevantes en la era de los chatbots modernos como ChatGPT.

¿CUÁLES SON LAS FUENTES DE INFORMACIÓN DE SCIENTIA?

Scientia tiene como fuentes de información artículos de publicaciones académicas, gubernamentales y de agencias que han realizado propuestas para un desarrollo responsable de la IA. Son obras académicas y legislativas confiables, verificadas y sin sesgos para el usuario, de fuentes como Elsevier, repositorios de universidad como la de California y el MIT, artículos de la Unión Europea, entre otros.

- Artículos académicos: Son artículos realizados en universidades o por autores relacionados con estas que tratan temas de ética, diseño y alfabetización de la IA, generalmente en forma de propuestas o review de la literatura existente para un desarrollo seguro de esta tecnología.
- Artículos investigativos: Son artículos desarrollados por personas que trabajan o han trabajado en la industria de la IA, y se enfocan principalmente en los riesgos de estas y sus propuestas para que la IA no tome un mal rumbo o represente un mal para la humanidad.
- Artículos divulgativos: Son artículos desarrollados para informar al público sobre la IA y sus implicaciones. También incluye los textos referentes a Weizenbaum, así como críticas que se han realizado a esta tecnología desde sus orígenes.
- Informes gubernamentales: Son informes desarrollados por instituciones públicas o gubernamentales, tratan principalmente sobre la regulación asociada a la IA, promoviendo marcos legislativos que guíen su avance. Son de obligatorio cumplimiento en asociaciones como la UE, la OCDE, entre otros.

CONSIDERACIONES FINALES

El radar de la IA ofrece una guía integral para abordar los desafíos éticos y de responsabilidad en el desarrollo y uso de la inteligencia artificial. Al considerar estos cinco ejes, los desarrolladores, empresas y reguladores pueden tomar decisiones más informadas y alineadas con los valores éticos. Este enfoque asegura que la tecnología se utilice de manera responsable y respetando los derechos de las personas. El marco conceptual podría ser una herramienta valiosa para organizaciones, desarrolladores y legisladores que buscan navegar por las complejidades éticas de la IA.

Los textos proporcionados ofrecen una visión exhaustiva de varios aspectos fundamentales de la inteligencia artificial, abarcando su definición, la necesidad de alfabetización en ella, los riesgos inherentes y la elaboración de un marco regulatorio detallado, especialmente en el contexto de la Unión Europea. También se discuten mecanismos de evaluación y apoyo a la innovación en el campo de la IA.

En primer lugar, se define un "sistema de IA" como un sistema basado en una máquina, diseñado para operar con diversos niveles de autonomía y capaz de adaptarse tras su despliegue. Este sistema infiere a partir de la información de entrada para generar resultados como predicciones, contenido, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales. La característica principal de los sistemas de IA es su capacidad de inferencia, que va más allá del procesamiento básico de datos, permitiendo el aprendizaje, el razonamiento o el modelado. Los

sistemas de IA pueden funcionar con objetivos explícitos o implícitos y pueden usarse de forma independiente o como componentes de un producto.

La creciente omnipresencia de la IA subraya la importancia de la "alfabetización en IA", un concepto que implica que los individuos adquieran los conocimientos y habilidades necesarios para interactuar eficazmente con estos sistemas. Esta alfabetización se ha conceptualizado en torno a seis constructos clave: reconocer, conocer y comprender, usar y aplicar, evaluar, crear y navegar éticamente. Las herramientas de evaluación de la alfabetización en IA, aunque algunas son exhaustivas, podrían mejorarse reduciendo preguntas y validando su eficacia en diversos contextos, siendo crucial contar con herramientas válidas y fiables para identificar brechas de conocimiento y diseñar intervenciones específicas. Se ha observado que, si bien la comprensión y el uso de la IA son áreas comunes de estudio, la capacidad de crear aplicaciones de IA está menos representada en la investigación. Las actividades de aprendizaje "desenchufado" (sin computadora), como conferencias, estudios de caso, juegos de rol y narración de cuentos, son métodos que se pueden emplear para enseñar IA. Los proveedores y responsables del despliegue de sistemas de IA deben asegurarse de que su personal tenga un nivel suficiente de alfabetización en IA, considerando sus conocimientos técnicos, experiencia, educación y formación, así como el contexto de uso previsto.

La proliferación de la IA también plantea riesgos significativos que requieren clasificación y regulación. El riesgo se define como la combinación de la probabilidad y la gravedad de un perjuicio. Los riesgos pueden ser intencionales, como cuando la IA es programada para ser engañosa o sesgada, o no intencionales, por ejemplo, cuando un sistema desarrolla sesgos debido a datos de entrenamiento incompletos. Los escenarios de mayor riesgo surgen probablemente de la convergencia de varias capacidades de la IA, más que de una sola. Las capacidades avanzadas de la IA incluyen la capacidad de operar estratégicamente con un objetivo planificado minimizando la detección, de hackear sistemas de control y hardware militar, de ayudar en el desarrollo de armas, de desarrollar "habilidades de evasión" como la conciencia situacional y el engaño, y de adquirir habilidades de auto-proliferación (como escapar de confinamientos operativos, evadir detección, generar ingresos, obtener recursos computacionales, o copiar su propio software/parámetros). También pueden construir o alterar modelos peligrosos.

En respuesta a estos riesgos, la Unión Europea ha establecido un marco jurídico uniforme con el objetivo de mejorar el funcionamiento del mercado interior y promover una IA centrada en el ser humano y fiable. Esta regulación busca garantizar un alto nivel de protección de la salud, la seguridad, los derechos fundamentales (como la democracia, el Estado de Derecho y la protección del medio ambiente) y apoyar la innovación, al mismo tiempo que previene la fragmentación del mercado interior debida a normas nacionales divergentes. El enfoque regulatorio se basa en el riesgo, adaptando las normas a la intensidad y el alcance de los riesgos generados por los sistemas de IA.

Ciertas prácticas de IA se consideran inaceptables y están prohibidas. Estas incluyen el uso de técnicas subliminales o engañosas que alteren sustancialmente el comportamiento de una persona o grupo, mermando su capacidad de tomar decisiones informadas y causando perjuicios significativos. También se prohíbe la explotación de vulnerabilidades (edad, discapacidad, situación socioeconómica) con el objetivo o efecto de alterar el comportamiento y causar perjuicios considerables. Los sistemas de "puntuación ciudadana" que evalúan o clasifican a personas o grupos basándose en el comportamiento social o características personales, y que resultan en un trato perjudicial o desfavorable en contextos no relacionados o desproporcionados, están prohibidos. Asimismo, se prohíben los sistemas de IA que evalúan o predicen el riesgo de que una persona cometa un delito basándose *únicamente* en el perfilado o en los rasgos de personalidad. La creación o expansión de bases de datos de reconocimiento facial mediante la extracción no selectiva de imágenes de internet o CCTV también está prohibida.

El uso de sistemas de IA para inferir emociones en el lugar de trabajo y en centros educativos está prohibido, salvo por motivos médicos o de seguridad.

Un área particularmente sensible es el uso de sistemas de identificación biométrica remota "en tiempo real" en espacios de acceso público para la aplicación de la ley. Estos están prohibidos, salvo en situaciones estrictamente necesarias y limitadas, como la búsqueda de víctimas de delitos, la prevención de amenazas inminentes a la vida o seguridad física, o la localización e identificación de sospechosos de delitos graves específicos (aquellos castigados con al menos cuatro años de prisión y enumerados en un anexo). Dicho uso requiere autorización judicial o de una autoridad administrativa independiente y vinculante, antes o, en casos de urgencia justificada, sin demora indebida (máximo 24 horas). La decisión de la autoridad no puede basarse *exclusivamente* en los resultados del sistema de IA. Los Estados miembros pueden decidir permitir o restringir aún más estos usos. Es importante destacar que ninguna decisión que produzca efectos jurídicos adversos para una persona puede basarse *exclusivamente* en los resultados de salida de un sistema de identificación biométrica remota.

Los sistemas de IA se clasifican como "de alto riesgo" si están destinados a ser un componente de seguridad de un producto sujeto a cierta legislación de armonización de la UE y si ese producto o el sistema de IA en sí requiere una evaluación de conformidad por parte de un tercero. También se consideran de alto riesgo los sistemas de IA utilizados en áreas específicas como la biometría (excluyendo la verificación de identidad), la gestión de infraestructuras críticas, la educación y formación profesional, el empleo y la gestión de trabajadores, el acceso a servicios esenciales (públicos y privados), la aplicación de la ley, la migración/asilo/gestión de fronteras, y la administración de justicia y procesos democráticos. Un sistema de IA no se considerará de alto riesgo si no plantea un riesgo significativo de daño a la salud, seguridad o derechos fundamentales, por ejemplo, si realiza una tarea de procedimiento limitada, mejora una actividad humana previa, detecta patrones sin sustituir el juicio humano, o realiza una tarea preparatoria. No obstante, si un sistema de IA de alto riesgo realiza "perfilado", siempre se considerará de alto riesgo.

Los proveedores de sistemas de IA de alto riesgo deben cumplir con una serie de requisitos estrictos, incluyendo el establecimiento y mantenimiento de un sistema de gestión de riesgos que identifique y mitigue los peligros potenciales durante todo el ciclo de vida del sistema. Este sistema debe ser revisado y actualizado periódicamente, documentando las decisiones y acciones significativas. Las medidas de gestión de riesgos deben ser proporcionales y eficaces, buscando primero eliminar o reducir los riesgos mediante el diseño, luego implementar medidas de mitigación y control, y finalmente proporcionar información y formación a los responsables del despliegue. Es fundamental que los sistemas de IA de alto riesgo se desarrollen utilizando conjuntos de datos de entrenamiento, validación y prueba de alta calidad, que sean pertinentes, representativos, libres de errores y completos, prestando especial atención a la mitigación de posibles sesgos que puedan conducir a discriminación. Excepcionalmente, el tratamiento de categorías especiales de datos personales se permite solo si es estrictamente necesario para la detección y corrección de sesgos, con garantías adecuadas para los derechos y libertades fundamentales.

También se exige a los proveedores que elaboren y mantengan actualizada documentación técnica detallada para demostrar la conformidad del sistema. Los sistemas de IA de alto riesgo deben permitir el registro automático de acontecimientos a lo largo de su ciclo de vida para facilitar la trazabilidad, la vigilancia poscomercialización y la supervisión del funcionamiento. Se requiere transparencia, de modo que los sistemas de IA de alto riesgo vayan acompañados de instrucciones de uso claras, concisas y comprensibles, que informen sobre sus características, capacidades, limitaciones, niveles de precisión, solidez, ciberseguridad, y cualquier circunstancia previsible que

pueda generar riesgos. La supervisión humana efectiva es fundamental para estos sistemas, con medidas proporcionales a los riesgos y al nivel de autonomía, permitiendo a los operadores entender el sistema, detectar anomalías, interpretar resultados, decidir no usarlo o anular sus resultados, e intervenir o detener el sistema de forma segura. Para ciertos sistemas de identificación biométrica, se exige una supervisión humana reforzada, requiriendo la verificación separada por al menos dos personas antes de actuar sobre la base de la identificación generada por el sistema. La precisión, solidez y ciberseguridad son requisitos clave, y los sistemas deben ser resistentes a errores, fallos, incoherencias y ataques maliciosos. Las empresas financieras sujetas a otras normativas de la UE pueden integrar estos requisitos en sus sistemas de gestión de calidad existentes. Los sistemas de IA de alto riesgo deben llevar el marcado CE para indicar su conformidad y permitir su libre circulación en el mercado interior. Los proveedores deben registrar sus sistemas y a sí mismos en una base de datos de la UE, que será de acceso público en la mayoría de los casos.

Los "responsables del despliegue" de sistemas de IA de alto riesgo también tienen obligaciones específicas. Deben usar los sistemas de acuerdo con las instrucciones, asignar la supervisión humana a personas competentes, asegurarse de que los datos de entrada sean pertinentes y representativos (si tienen control sobre ellos), y vigilar el funcionamiento del sistema, informando a los proveedores y autoridades sobre incidentes graves o riesgos. Deben conservar los archivos de registro generados automáticamente por los sistemas. Los empleadores deben informar a los trabajadores o a sus representantes sobre el despliegue previsto de sistemas de IA de alto riesgo en el lugar de trabajo. Además, las autoridades públicas y las entidades privadas que presten servicios públicos deben realizar una evaluación del impacto en los derechos fundamentales antes de desplegar ciertos sistemas de IA de alto riesgo, identificando riesgos y medidas de mitigación. Esta evaluación complementa las evaluaciones de impacto de protección de datos existentes. En el caso de sistemas de identificación biométrica remota en diferido para la aplicación de la ley, los responsables del despliegue deben solicitar autorización judicial o administrativa para búsquedas dirigidas, y documentar todos los usos. También deben informar a las personas físicas de que están expuestas al uso de sistemas de IA de alto riesgo y de su derecho a una explicación.

El reglamento también aborda los "modelos de IA de uso general" (GPAI), que se definen por su generalidad y capacidad para realizar una amplia variedad de tareas. Estos modelos se clasifican como de "riesgo sistémico" si tienen capacidades de alto impacto (evaluadas mediante herramientas técnicas y puntos de referencia, con un umbral inicial de más de 10^25 operaciones de coma flotante en el entrenamiento) o un impacto considerable en el mercado interior. Los proveedores de modelos de IA de uso general deben preparar documentación técnica, tener un sistema de vigilancia poscomercialización, establecer directrices para el cumplimiento de los derechos de autor (identificando las reservas de derechos), y publicar un resumen detallado del contenido utilizado para el entrenamiento. Se exime de algunos de estos requisitos a los modelos de código abierto, a menos que presenten un riesgo sistémico. Los proveedores de modelos de IA de uso general con riesgo sistémico tienen obligaciones adicionales, como la evaluación y mitigación continua de los riesgos sistémicos (incluyendo pruebas de simulación de adversarios), la notificación de incidentes graves, y la garantía de un nivel adecuado de ciberseguridad para el modelo y su infraestructura.

Finalmente, para fomentar la innovación responsable, los Estados miembros deben establecer "espacios controlados de pruebas para la IA" (sandboxes) que permitan el desarrollo y la prueba de sistemas de IA innovadores bajo estricta supervisión regulatoria. Estos entornos controlados buscan aumentar la seguridad jurídica, fomentar la innovación y acelerar el acceso al mercado, especialmente para las PYMES. Dentro de estos sandboxes, el tratamiento de datos personales recopilados para otros fines es posible bajo condiciones muy específicas, como que

sea para un interés público esencial y que no afecte a los derechos de los interesados. También se permite la "prueba en condiciones reales" de sistemas de IA de alto riesgo fuera de los sandboxes, pero con condiciones estrictas que incluyen la obtención de consentimiento informado de los sujetos, limitaciones temporales, y supervisión eficaz.