Inteligencia artificial, Responsabilidad humana

Documento de posición de Triodos Bank sobre la IA ética

Triodos Bank

Alcance del documento

La inteligencia artificial (IA) ha ganado terreno progresivamente y ha suscitado debates acalorados, sobre todo desde el lanzamiento de Chat GPT en noviembre de 2022. En la actualidad se ha convertido en una cuestión empresarial de importancia máxima. Cada vez son más las aplicaciones y usos de los sistemas de IA para empresas y gobiernos y aunque las empresas especializadas en este ámbito se han convertido en el centro de atención entre quienes invierten, es evidente que hay riesgos considerables en torno a estas tecnologías. Es necesario conocer esos riesgos y la sociedad las empresas, los inversores e inversoras y cada persona individualmente los tengamos en cuenta.

Este documento pretende exponer la posición de Triodos Bank respecto a la IA desde el punto de vista de sus valores. Para ello nos centramos principalmente en identificar los problemas y riesgos principales relacionados con los sistemas de IA. En el documento también formulamos principios de precaución respecto a los sistemas de IA, identificamos usos muy controvertidos de esta tecnología y analizamos el papel de las instituciones financieras para motivar a las empresas hacia prácticas responsables en el desarrollo y uso de la tecnología de IA.

Este documento de posición no clasifica los sistemas y tecnologías de IA ni analiza el impacto positivo actual o potencial de la tecnología de IA. Asimismo, se evita describir las implicaciones macroeconómicas de una adopción a gran escala de esta tecnología, así como sus efectos en el mercado laboral.

Índice

1. Introducción

"La IA no tiene nada de artificial. Está inspirada en las personas, creada por personas y, lo que es más importante, afecta a las personas. Es una herramienta poderosa que apenas empezamos a comprender, lo que supone una responsabilidad enorme".

Fei-Fei Li, co-directora del Stanford Institute for Human-Centered Artificial Intelligence.

La tecnología moldea nuestras sociedades y nuestras vidas desde los orígenes de la historia de la humanidad. La segunda mitad del siglo XX se vio marcada por la llegada de la informática y actualmente el desarrollo y adopción acelerados de los sistemas de inteligencia artificial (IA) marcan otro giro clave en la historia de la tecnología y, posiblemente, de la humanidad.

El lanzamiento de ChatGPT en 2022 marcó el inicio de un debate público acalorado en torno a la IA. La tecnología de la IA está considerada como un vector de transformación que cambia nuestras sociedades y economías y que ya da forma a nuestro futuro. Ya no se trata de si estamos a favor o en contra de la IA, sino de cómo asegurarnos de que la desarrollemos y utilicemos con criterio para que su poder y potencial generen efectos positivos.

Los sistemas de IA y sus usos han dejado de centrarse en demostrar teoremas[1] y jugar a las damas[2] y ahora se enfocan en desarrollar aplicaciones comerciales de gran impacto. Actualmente la IA genera un efecto profundo en la humanidad al estimular el progreso en muchas áreas de la ciencia y la economía a una velocidad sin precedentes. El aumento de la capacidad para recopilar datos ha permitido aprovechar grandes cantidades de información para desarrollar los sistemas de IA y ofrecer soluciones a problemas complejos en empresas, instituciones y en la sociedad civil. Muchos sectores utilizan la IA desde hace años, como es el caso de las ciencias de la vida para la obtención de imágenes médicas y el desarrollo de fármacos, la ciberseguridad para la detección de anomalías y fraudes, el diseño y la visualización en 3D, la visión por ordenador, y las ventas y el marketing para la elaboración de modelos predictivos.

Sin embargo, los avances conseguidos no han estado exentos de preocupaciones, y cada vez existe una mayor conciencia sobre los riesgos asociados a este rápido desarrollo tecnológico. Existen muchos ejemplos de sesgo algorítmico y, aunque no todos los algoritmos se basan en la IA, pueden tener consecuencias muy reales en las personas y la sociedad. Por ejemplo, los accidentes causados por vehículos de conducción autónoma han generado dudas. Entretanto, los contenidos generados por IA son cada vez más difíciles de diferenciar de los generados por seres humanos, mientras que la adopción de los sistemas de IA por parte de las empresas suscita preocupaciones por la sustitución de trabajadores.

Este documento expone la posición de Triodos Bank —desde la óptica de sus valores— en relación con la IA, con un enfoque preventivo. Se centra en los riesgos y problemas relacionados con los sistemas de IA. Para ello definimos principios de precaución respecto al desarrollo y al uso de sistemas de IA e identificamos casos de su uso que resultan muy controvertidos y que consideramos que deberían prohibirse. También destacamos el papel de las instituciones financieras para establecer los requisitos y expectativas de las empresas a las que financian y en las que invierten, así como en sus propias actividades y desde su ciudadanía corporativa. Concluimos con un llamamiento a la acción para que las instituciones financieras adopten una postura responsable respecto a la tecnología de IA.

Una nota importante: el hecho de que en este documento nos centremos en los aspectos críticos de la tecnología de IA no significa que Triodos Bank no crea en su potencial. Como banco con valores e inversor de impacto responsable nos interesan mucho los sistemas de IA que muestren un verdadero potencial de impacto positivo en cualquiera de los temas de transición que motivan nuestra estrategia de impacto. Sin embargo, creemos que esos impactos positivos no se derivan de la tecnología en sí, sino de cómo se utiliza, mientras que el impacto negativo puede que además dependa también de cómo se diseñe y se desarrolle. Una vez más nos corresponde a todos y todas sopesar los pros y los contras a través del desarrollo de un criterio (humano) óptimo.

1.1 Hablemos de inteligencia artificial

Las conversaciones sobre las ventajas y riesgos de la tecnología suelen desarrollarse sin una buena comprensión de la terminología y del contexto técnico en el que se enmarcan. Por lo tanto, es importante identificar los principales conceptos y definiciones y desmitificar el término inteligencia artificial (IA).

No existe una definición universalmente aceptada de IA. La UE y la OCDE[3] indican que un sistema de IA es *«un sistema basado en máquinas que, a través de objetivos explícitos o implícitos y a partir de los elementos de información que recibe, desarrolla la generación de unos resultados, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales. Los sistemas de IA varían en sus niveles de autonomía y adaptabilidad una vez desplegados».*

En pocas palabras, la **inteligencia artificial** es la capacidad de una máquina para realizar tareas comunmente asociadas a seres inteligentes. Actualmente, dentro de la tecnología de la IA el subconjunto más importante de técnicas utilizadas es el **aprendizaje automático**, que es la capacidad de un programa informático o de una máquina para aprender y realizar acciones sin que se le codifiquen explícitamente órdenes para hacerlo. Sin embargo, el término IA se utiliza de forma diferente según los grupos de personas que lo usen y, en el discurso general, lo que se denomina IA suele ser un conjunto de técnicas de la ciencia del aprendizaje automático denominado **aprendizaje profundo.** Esas técnicas permiten, en términos generales, simular artificialmente un proceso de aprendizaje humano a partir de datos no estructurados, como imágenes o archivos de audio. Las técnicas de aprendizaje profundo son responsables de la fama de algunos sistemas de IA como «cajas negras», mientras que otras técnicas de IA permiten explicar completamente los mecanismos que conducen al resultado de un modelo. Los modelos de la **IA generativa (GenAI)** están basados en técnicas de aprendizaje profundo que son capaces de generar contenidos escritos, visuales o de audio nuevos. ChatGPT es actualmente el ejemplo más conocido de ese tipo de IA.

Es importante señalar que no toda la IA es «superinteligente.» Al contrario, las aplicaciones actuales de IA se conocen como ***narrow AI*** o **«IA estrecha».** Esto significa que son sistemas diseñados para realizar un conjunto específico de tareas, como el software de filtrado de spam en el correo electrónico, los sistemas de traducción automática, los anuncios comerciales, los *chatbots* (incluido ChatGPT) y los vehículos de conducción autónoma. Todos esos sistemas operan dentro de los límites de un conjunto específico de objetivos. La perspectiva de la **inteligencia general artificial (AGI),** también llamada IA fuerte o cada vez con más frecuencia ***artificial superintelligence (ASI)*** o **superinteligencia artificial,** genera interrogantes y preocupaciones. Esos tipos de inteligencia artificial serían capaces de realizar tareas intelectuales de forma comparable a la de un ser humano y en el caso de la ASI más allá de la inteligencia humana. Aprendena adaptarse a situaciones nuevas y no se limitan a un conjunto específico de tareas.

Los sistemas de IA no son **robots** similares a los humanos. La robótica es una rama de la ingeniería mecánica que puede, aunque no necesariamente debe, utilizar técnicas de aprendizaje profundo (con resultados impresionantes). Sin embargo, no todos los robots están equipados con sistemas de IA, ni todos los sistemas de IA se mueven en el mundo físico. Además, no todos los modelos que utilizan datos están impulsados por la IA. La **ciencia de los datos** puede hacer uso de la IA, pero un simple conjunto de regresiones lineales —algo muy utilizado en la ciencia de datos— no es un sistema de IA.

Cuando hablamos de la IA es esencial ser conscientes de cómo el lenguaje determina nuestra forma de pensar y de entender las cosas. Por eso hemos optado conscientemente por referirnos a los «sistemas de IA» o «modelos de IA» como tales, en lugar de utilizar el término «IA» por sí solo. De ese modo se contribuye a desmitificar la inteligencia artificial, a evitar pensar en ella como una tecnología misteriosa y potencialmente sensible y a destacar que los seres humanos son los responsables últimos de cómo se diseña, se desarrolla y se utiliza la inteligencia artificial[4].

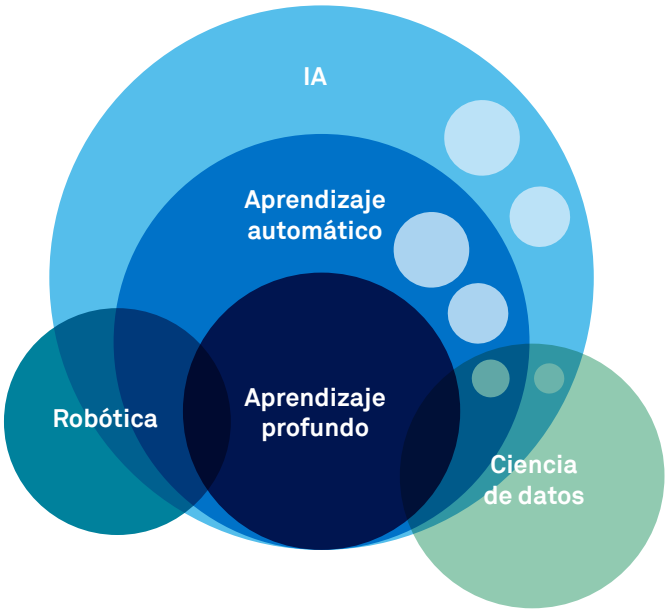

Imagen 1: Ilustración de interacción de diferentes disciplinas relacionadas con la IA.
Fuente: Elaboración propia de Andrew Ng, aprendizaje profundo de IA.

1.2 La regulación que viene

En relación con los avances tecnológicos, hasta hace poco los organismos reguladores se centraban sobre todo en los riesgos relacionados con el uso indebido de los datos personales. La difusión rápida de los sistemas de IA ha aumentado las preocupaciones, que ahora van más allá de la protección de datos. Las preocupaciones éticas y de seguridad relativas al desarrollo y uso de la tecnología de IA han motivado que varios gobiernos nacionales y organizaciones internacionales desarrollen y adopten directrices y marcos para una IA responsable. Desde marzo de 2024, algunos gobiernos nacionales e instituciones supranacionales también han desarrollado iniciativas regulatorias que abordan el desarrollo y el uso de la IA.

Los intentos de regular la IA se dirigen principalmente a prevenir y mitigar los riesgos relacionados con la equidad algorítmica, la transparencia y la supervisión humana[5]. En marzo de 2024, las instituciones de la UE aprobaron la Ley de Inteligencia Artificial de la UE (también llamada Ley de IA) que en el momento de publicar este documento se esperaba que entrera en vigor en junio de 2024. La normativa se aplica a los proveedores que comercializan sistemas de IA en el mercado de la UE o que los ponen en servicio en la UE por lo que, en la práctica, tiene un impacto global. Sin embargo, no se aplica a los sistemas de IA desarrollados en la UE y exportados fuera de sus fronteras. Algunos sectores quedan fuera de este documento regulatorio, entre ellos la aviación civil y la seguridad nacional. Al otro lado del Canal de la Mancha, el Reino Unido ha adoptado una postura más flexible, con un enfoque favorable a la innovación en la regulación de la IA, con la publicación un libro blanco en 2023 que proporciona un marco de gestión de riesgos al tiempo que se centra en apoyar la innovación.
En Estados Unidos hay varias iniciativas en marcha. En octubre de 2023 la presidencia estadounidense emitió una Orden Ejecutiva sobre Inteligencia Artificial Segura y Fiable, lo que indica la intención del gobierno de desarrollar normas sólidas para la seguridad de la IA y en otros ámbitos. Por su parte, China ha optado por una regulación más específica de la IA, con una normativa sobre algoritmos de recomendación[1] en 2021, normas sobre contenidos generados de forma sintética en 2022 y una regulación sobre IA generativa en 2023. En Canadá, Japón, Corea del Sur, Australia, Singapur e India también se realizan esfuerzos notables en este sentido. En general, los países de África y América Latina son menos activos en la materia.

A escala mundial, la ONU ha creado un órgano consultivo sobre IA para hacer recomendaciones sobre la gobernanza internacional en este ámbito[6]. También surge una red mundial creciente de Institutos de Seguridad de la IA —con el compromiso de una colaboración estrecha de varias partes, incluidos EE.UU. y el Reino Unido— en la que los modelos de IA de última generación se someterían a pruebas rigurosas antes de lanzarse al público[7]. Entretanto, los países del G7 han acordado un código de conducta voluntario[8] para las organizaciones que desarrollan sistemas avanzados de IA y que complementa las iniciativas regulatorias.

Varios llamamientos a la regulación –pero con prioridades diferentes

Aunque la industria de la IA tiene una postura muy positiva sobre sus beneficios potenciales para la humanidad, agentes importantes del sector han reclamado una regulación y unos sistemas de gobernanza sólidos, sobre todo con respecto a la denominada inteligencia general artificial (artificial general intelligence, AGI) y la superinteligencia. Las empresas dedicadas a la IA son especialmente elocuentes sobre las amenazas existenciales vinculadas a su rápido desarrollo y han mostrado su preocupación por la capacidad para controlar ese ritmo y la dirección del desarrollo de la tecnología, sobre todo en vista de la presión competitiva y accionarial a la que se enfrentan estas empresas.

Por el contrario, al exigir una mayor regulación, las organizaciones de la sociedad civil y las y los investigadores académicos ponen de relieve los problemas existentes con el desarrollo y el uso de la IA, en particular el sesgo algorítmico y las vulneraciones de derechos humanos y libertades universales. En la UE, una coalición de organizaciones de la sociedad civil liderada por la Red Europea de Derechos Digitales (EDRi) pide desde hace tiempo que la Ley de Inteligencia Artificial de la UE proteja y promueva los derechos humanos, con especial atención a los grupos marginados. De ese llamamiento se han hecho eco universidades de renombre y muchas organizaciones independientes que advierten también del riesgo de que existan lagunas en la legislación que permitan a las empresas tecnológicas autorregularse.

[1] Se trata de algoritmos que proporcionan a cada persona usuaria sugerencias personalizadas que se consideran más pertinentes. Son el tipo de IA más desplegado en Internet.

2. Cuestiones y riesgos éticos y de sostenibilidad

Muchas de las cuestiones éticas y de sostenibilidad que surgen en relación con la tecnología de la IA no son nuevas, y tampoco son específicas. Por ejemplo, los debates sobre el sesgo algorítmico y la energía necesaria para el funcionamiento de los centros de datos se producen desde antes de que estos sistemas se adoptaran a la escala y el ritmo actuales. Sin embargo, la omnipresencia de la tecnología de IA les da una dimensión nueva . Creemos que esto se debe a una característica intrínseca de la tecnología que multiplica los problemas (amplificación), así como a la tendencia de las personas usuarias a confiar demasiado en la tecnología (dependencia excesiva) y a la posibilidad de trasladar las responsabilidades a las máquinas (externalización moral).

Creemos que estos tres aspectos —tanto en la tecnología de la IA como en nuestras actitudes hacia ella— pueden ofrecernos una lente con la que analizar los riesgos y problemas relacionados con la IA. Esos aspectos pueden agruparse en categorías más clásicas, como uso de recursos (o riesgos medioambientales), derechos humanos y derechos fundamentales, y ética empresarial.

Raíces de las cuestiones éticas y de sostenibilidad relacionadas con la IA

- **Amplificación.** Los sistemas de IA tienen una capacidad sin precedentes para operar a escala. Esto significa que también tienen un potencial enorme para amplificar los problemas no abordados y exacerbar los prejuicios y las amenazas relacionadas con nuestro tejido social. Al mismo tiempo, el uso de la tecnología de la IA puede desencadenar un cambio en el uso de los recursos porque la innovación tecnológica aumenta la necesidad de recursos naturales frente a la mano de obra humana.
- **Exceso de confianza.** Los sistemas de IA se tratan a menudo como fuentes de conocimiento (y a veces incluso de sabiduría) altamente racionales y fiables. Sus resultados se dejan a menudo sin escrutar y sin cuestionar, como si fueran oráculos. Las personas usuarias corren el riesgo de confiar ciegamente en los resultados de los sistemas de IA y esperar que ofrezcan soluciones a prácticamente cualquier problema (incluso a aquellos que pueden resolverse en un plazo y con un uso de tecnologías más primitivas).
- **Externalización moral.** El exceso de confianza se explica en parte por la tendencia a referirse a los sistemas de IA como seres sensibles, al hablar por ejemplo de “IA racista” o de “máquinas xenófobas”. Si no utilizamos el lenguaje de forma adecuada, podemos trasladar la responsabilidad sobre los prejuicios codificados y sobre su funcionamiento a los propios productos de la IA automáticamente. Esto puede “absolver” a quienes crean y utilizan los sistemas de IA de las obligaciones morales asociadas a la difusión de esta tecnología.

2.1. Uso de recursos

La huella de carbono de los modelos de IA lleva es objeto de escrutinio público desde hace mucho tiempo. Aunque todavía no se dispone de estimaciones ampliamente reconocidas sobre los costes ambientales de la IA, a continuación, exponemos tres áreas principales de preocupación sobre grandes costes ambientales en el desarrollo y despliegue de la IA.

- El **consumo energético** de los grandes centros de datos vinculados al entrenamiento de los sistemas de IAy la energía necesaria para desplegar esos sistemas son difíciles de cuantificar. Esto se debe en parte a la falta de información por parte de las empresas que suministran la tecnología. Sin embargo, un estudio reciente calcula que para 2027el consumo energético de los sistemas de IA será probablemente equivalente al de un estado de las dimensiones de los Países Bajos, por ejemplo[9]. Otros estudios predicen que para 2030 la IA podría representar entre el 3 % y el 4 % de la demanda mundial de energía[10]. Con el rápido aumento del uso de esos sistemas su huella de carbono de va a tener un papel cada vez mayor en el deterioro del medioambiente.

- El **consumo y la extracción de agua** para la tecnología de IA a menudo no se tiene en cuenta y se relaciona sobre todo con el agua utilizada en los centros de datos para generar electricidad y refrigerar los servidores. Sin embargo, un estudio reciente[11] ha descrito y estimado el consumo de agua relacionado con la IA con una réplica del enfoque de alcance utilizado para estimar las huellas de carbono. Ese estudio no solo tiene en cuenta la refrigeración de los servidores *in situ* (alcance 1), sino el agua para la generación de electricidad (alcance 2) y de la cadena de suministro para la fabricación de servidores (alcance 3). Los resultados dibujan un panorama muy negativo en cuanto al uso de agua. La dependencia de los modelos de IA también plantea problemas en cuanto a la extracción de recursos hídricos porque

el agua dulce es un recurso finito y se agota y se contamina más rápido de lo que se repone. Esto hará que surja la competencia entre sectores de la economía por el uso del agua en las regiones y momentos en los que se encuentren.

- **Materias primas y residuos electrónicos.** La cuestión del uso de materias primas y residuos electrónicos no es específica de la IA, sino que se aplica a la tecnología en su conjunto. En 2019, se producían más de 50 millones de toneladas métricas de residuos electrónicos al año[12]. Las aplicaciones de IA pueden ser potencialmente útiles y eficaces para reducir la cantidad de materia prima utilizada en los procesos industriales, desde la producción hasta la gestión de residuos. Sin embargo, su adopción en la industria podría llevar tiempo. El desarrollo rápido de los sistemas de IA también significa que su *hardware* puede quedar obsoleto rápidamente, lo que contribuye al problema de los residuos electrónicos que si no se reciclan y se tratan adecuadamente liberan al medioambiente sustancias químicas peligrosas como plomo, mercurio y cadmio.

Los problemas medioambientales relacionados con la IA están relacionados sobre todo con la fase de desarrollo de los modelos, en la que se utilizan grandes cantidades de recursos y energía durante las pruebas. El despliegue de los sistemas de IA también es un factor a tener en cuenta, pero en menor medida. Se trata de un inconveniente de la propia tecnología más que el resultado de un enfoque poco ético respecto al desarrollo de la IA. Sin embargo, además de los aspectos técnicos del uso de los recursos y la energía, debemos reconocer nuestra gran dependencia —a veces excesiva— de la tecnología digital, incluida la IA, como uno de los motores de su consumo. Por lo tanto, esto se añade a las cuestiones ambientales descritas anteriormente.

2.2 Derechos humanos y derechos fundamentales

Los derechos humanos están consagrados en la Declaración de los Derechos Humanos de la ONU[13], así como en la Carta de los Derechos Fundamentales de la UE incluida en el Tratado de Lisboa[14] y en el Pacto Internacional de Derechos Civiles y Políticos[15]. El desarrollo y uso generalizado de los sistemas de IA plantea preguntas sobre la vulneración de varios derechos humanos y fundamentales tanto de las personas como de las entidades colectivas.

- El **derecho al respeto de la vida privada y la intimidad** es un derecho humano fundamental e implica que la información personal, incluidos los registros oficiales, fotografías, cartas, diarios e historiales médicos, deben conservarse de forma segura y solo compartirse con permiso de la persona interesada. Los algoritmos de IA se basan y se prueban con datos que pueden incluir algunos personales. La privacidad de los datos de un usuario o usuaria se vulnera cuando se recogen y utilizan sin su conocimiento, por ejemplo, con fines de publicidad dirigida. Aún más preocupante es la presencia de sistemas de reconocimiento biométrico, como los de reconocimiento facial utilizados en espacios públicos. Esto supone una amenaza muy importante para la privacidad porque se pueden controlar los movimientos y el comportamiento de las personas en tiempo real sin su consentimiento expreso. Por último, aunque la IA desempeña ahora un papel fundamental en la ciberseguridad, los sistemas de IA también pueden utilizarse para lanzar ciberataques. Las y los ciberdelincuentes apuntan cada vez más a los sistemas de IA, lo que supone una amenaza creciente para la privacidad y la seguridad de los datos de particulares y empresas.

- La **no discriminación** es de importancia vital en el desarrollo de los sistemas de IA, que pueden incorporar sesgos en cualquier fase del ciclo de vida de los modelos[16], desde la recopilación de datos y el desarrollo de modelos hasta su despliegue, el seguimiento y la integración de información. Los modelos de IA se nutren con grandes cantidades de datos que pueden estar sesgados, lo que supone la amplificación y perpetuación de estereotipos. Los sistemas de IA resultantes pueden discriminar a personas infrarrepresentadas o sobrerrepresentadas en la muestra utilizada para el desarrollo. Además, los equipos de desarrolladores pueden tener prejuicios inconscientes que se reflejan en los datos y que se pasan por alto. Este tipo de problemas se han identificado en muchos campos, como la sanidad[17] y la banca(riesgo de crédito)[18]. Se ha demostrado que las predicciones algorítmicas discriminan a los grupos minoritarios y tienen un sesgo racial grave por razones que no pueden atribuirse únicamente a conjuntos de datos sesgados.

- El **derecho a la vida, la libertad y la seguridad personal** se ve cada vez más amenazado a medida que se generalizan los sistemas de armas autónomas[19]. Se espera que para 2030 los vehículos robóticos teledirigidos puedan participar en enfrentamientos bélicos junto con las tropas convencionales, al tiempo que los vehículos de combate remotos y totalmente autónomos incorporarían sus propios sistemas de IA[20]. En un caso extremo de externalización moral, esos tanques automatizados, robots asesinos y drones kamikaze podrían tomar decisiones de vida o muerte sin que fuera necesaria la intervención de un ser humano. Además de la seguridad física, cuestiones como la ciberdelincuencia y las amenazas a la propiedad y la identidad digital de las personas son ya una realidad. La accesibilidad y disponibilidad pública de sistemas sofisticados de IA plantean riesgos de seguridad graves, tanto para las personas como

para la sociedad en su conjunto. Las y los ciberdelincuentes pueden utilizar la IA para generar software malicioso, automatizar ataques y aumentar la eficacia de las estafas mediante aplicaciones *deep fake*[21].

- El **derecho a la libertad de opinión y de expresión** (así como el derecho a la libertad de pensamiento, conciencia y religión, y de reunión pacífica) se puede vulnerar fácilmente cuando se somete a las ciudadanas y ciudadanos a una vigilancia monitorizada por la IA y que posteriormente utilizan por las fuerzas del orden. Sin salvaguardas normativas sólidas, la ciudadanía corre el riesgo de ser perseguida por ejercer su derecho a criticar a un régimen gobernante o a participar en manifestaciones políticas o reuniones culturales y religiosas. Además, pueden utilizarse herramientas de IA para censurar a los medios de comunicación y a las y los periodistas independientes. La concentración de poder en manos de unas pocas plataformas privadas online aumenta el riesgo de que la información sea gobernada ilegalmente[22].

- El **derecho a la protección de la propiedad intelectual** es un derecho humano reconocido en la Declaración (artículo 27.2) y un derecho fundamental recogido en la Carta de la UE (artículo 17) y también es esencial para la innovación. Asimismo, pueden producirse infracciones de los derechos de propiedad intelectual si los sistemas de IA se entrenan o funcionan con datos procedentes de Internet como textos e imágenes sin comprobar si están protegidos o son privados. Los sistemas de IA pueden generar resultados que reproduzcan parcialmente textos sin proporcionar las fuentes o generar imágenes muy parecidas al material original[23].

- Los **derechos laborales** son un subconjunto de los derechos humanos (artículo 23 de la Declaración) y la aparición de sistemas de IA que podrían utilizarse para tareas anteriormente realizadas por humanos plantea riesgos sociales a gran escala[24]. La IA puede automatizar procesos, generar textos y presentaciones y realizar análisis con una eficiencia superior a la de la mayoría de las personas. Sin embargo, aunque esos riesgos plantean interrogantes para el futuro del trabajo, algunos derechos laborales ya están amenazados o vulnerados en el proceso de desarrollo de la IA. El entrenamiento de los sistemas actuales de IA requiere la aportación humana de quienes etiquetan los datos, lo que, al parecer, ha dado lugar a una industria de millones de personas trabajadoras en todo el mundo que realizan tareas repetitivas en condiciones laborales precarias y con salarios ínfimos. Esas personas suelen estar expuestas a contenidos violentos o perturbadores[25].

En resumen, algunas tecnologías de IA, como el reconocimiento facial en vivo y la identificación y clasificación biométricas, suscitan preocupaciones específicas tanto por su uso potencial para la vigilancia masiva como por su fiabilidad desde una perspectiva de no discriminación. Esto amplifica el riesgo de discriminación estructural y de limitación de las libertades personales. Los sistemas de armamento autónomos pueden generar prácticas bélicas, de defensa y de seguridad nuevas al externalizar en las máquinas la responsabilidad de tomar decisiones de vida o muerte, lo que constituye un ejemplo claro de externalización moral. Por último, la cuestión de los algoritmos sesgados es una preocupación que puede afectar a cualquier uso de la IA, con implicaciones directas o indirectas para el acceso de las personas a un nivel de vida adecuado.

2.3. Ética empresarial

La ética empresarial es la aplicación de valores éticos y principios morales a la forma en que las empresas y l las personas participan en las actividades de negocio. La ética empresarial abarca una gama amplia de temas, incluido el gobierno corporativo, y suele venir determinada por políticas y procedimientos. Los principios de ética empresarial y las prácticas que se derivan de ellos contribuyen a generar confianza en las empresas y a menudo preceden y complementan a la regulación. En el ciclo de vida de los sistemas de IA existen varias etapas —desde la conceptualización, el desarrollo o el diseño hasta el despliegue y el uso— en las que se debe tener en cuenta la ética empresarial.

- La **falta de transparencia y aplicabilidad** de los sistemas de IA es una de las deficiencias más discutidas de la tecnología y la causa fundamental de varias de las cuestiones presentadas en este capítulo. Aunque es difícil establecer normas claras para la transparencia de los sistemas de IA[26] es importante que los resultados de los sistemas sean explicables. Eso significa que los conjuntos de datos subyacentes deben estar disponibles para su revisión y que los algoritmos deben poder reproducirse para que puedan detectarse los errores y fallos. Por ejemplo, si se sospecha que los resultados están sesgados, debe ser posible comprobar si la muestra de datos subyacente también lo está de alguna manera o en qué parte del sistema radica el problema. Es fundamental que exista transparencia en la ética empresarial para generar confianza y la debida rendición de cuentas tanto en la empresa como en sus productos.

- La **falta de una rendición de cuentas clara** constituye otro motivo de preocupación. Resulta difícil responsabilizar a los agentes de la cadena de producción de los sistemas de IA y de los productos impulsados por ella porque los propios resultados son difíciles de explicar. Por eso es tan importante que exista una buena gobernanza con respecto a los productos, que es la supervisión activa del diseño, del desarrollo, del cumplimiento, de la gestión de riesgos y de la protección de la confian-

za. Un ejemplo claro son los vehículos de conducción autónoma, que utilizan software de IA y sensores para desplazarse entre destinos sin interferencia humana. Aunque esa tecnología puede revolucionar los desplazamientos y reducir los errores humanos que provocan accidentes, el sistema de IA también puede fallar a la hora de identificar riesgos o peatones en la calle. Esos ejemplos también plantean cuestiones sobre la responsabilidad, la seguridad de los productos y la concienciación de quienes los utilizan.

- Las **prácticas manipuladoras de marketing** se ven enormemente favorecidas por los modelos de IA, que se utiliza ampliamente en campañas dirigidas y basadas en el historial del navegador y en las cookies de las personas usuarias, así como en sus historiales de compras online. Las empresas han de garantizar que sus prácticas de marketing sean justas para las y los clientes. Sin embargo, existe una línea muy fina entre personalizar la experiencia del cliente y manipularla y las empresas tienen que dirimir esa cuestión.

- La adopción de **mecanismos que inducen a la adicción** ha sido objeto de gran escrutinio desde la aparición de las redes sociales y estos mecanismos pueden verse exacerbados por los últimos avances de la IA. La dependencia creciente de las redes y herramientas controladas por máquinas puede tener un impacto negativo en las capacidades cognitivas y sociales de las personas. Sin embargo, se trata de una práctica bastante habitual para muchas empresas que se centran en interactuar con las personas usuarias y desarrollan de manera expresa sistemas que pueden provocar adicciones graves.

En resumen, los sistemas de IA plantean preocupaciones importantes en cuanto a la transparencia del diseño y el desarrollo y la capacidad de las empresas para garantizar una rendición de cuentas adecuada y para asumir una responsabilidad plena durante el ciclo de vida de la IA. Además, los sistemas de IA pueden generar características altamente adictivas y utilizarse con intenciones manipuladoras.

3. Adoptar una postura con respecto a la IA responsable

Como ya se ha señalado, el desarrollo o uso irresponsable o sin principios de los sistemas de IA puede entrañar riesgos considerables para los derechos humanos y laborales, la ética empresarial y la seguridad individual y colectiva. Esos riesgos no pueden pasarse por alto. La tecnología de la IA es una herramienta que no es intrínsecamente buena o mala. Puede tener efectos positivos cuando se utiliza de forma consciente, pero también puede generar efectos negativos en función de cómo se diseñe, desarrolle e implante.

Por eso es esencial definir algunos principios generales que sienten las bases para un desarrollo y un uso responsables de la tecnología de la IA de una forma socialmente amplia. También es fundamental establecer límites sobre lo que son prácticas y usos altamente controvertidos de los sistemas de IA.

3.1 Principios para un uso y desarrollo responsables de la tecnología de IA

Varios organismos gubernamentales, organizaciones internacionales y de la sociedad civil, así como algunas empresas han definido principios para una IA fiable. Por su parte, la UE ha establecido unas directrices éticas en ese mismo sentido. En el cuadro de texto siguiente se incluyen esos principios rectores. Aunque la redacción exacta podría cambiar ligeramente, la mayoría de las entidades han enumerado principios similares[2] con los que estamos de acuerdo en líneas generales.
En nuestras propias palabras, estos son los principios rectores que creemos que debe cumplir la tecnología de IA para ser realmente beneficiosa.

IA centrada en la humanidad. Creemos que los sistemas de IA y la tecnología en general deben poner la dignidad humana en el centro. No solo en el ser humano, sino en la humanidad. Esto quiere decir que los sistemas de IA deben adaptarse a nuestras necesidades y preferencias como personas tener en cuenta el bienestar en su sentido más amplio cuando se diseñan y se desarrollan. En otras palabras, no basta con que la tecnología esté diseñada para que las personas puedan entenderla y utilizarla para satisfacer sus necesidades individuales. También debe diseñarse, desarrollarse y fabricarse de forma que satisfaga nuestras necesidades y libertades colectivas. Eso requiere un enfoque consciente de los recursos empleados. La tecnología debe ayudar a la humanidad a conseguir unas sociedades prósperas y sanas y que las personas vivamos en un planeta saludable y acogedor.

Principios rectores

Según las directrices éticas de la UE para una IA fiable, los sistemas de IA deben ser:

- **Legales.** Deben respetar todas las leyes y normativas aplicables
- **Éticos.** deben respetar principios y valores.
- **Robustos** desde un punto de vista técnico y tener en cuenta el entorno social.

IA fiable. Es esencial que los sistemas de IA se diseñen y desarrollen con los estándares de seguridad técnica y robustez más altos. Deben probarse adecuadamente a lo largo del tiempo antes de ponerlos en circulación. Los datos utilizados deben ser de alta calidad y su acceso legítimo. Los datos personales y los datos en general deben recopilarse, almacenarse y utilizarse de forma responsable, de acuerdo con las normas de privacidad y seguridad. Debe haber transparencia siempre que interactuemos con máquinas en lugar de con otras personas y debe explicarse adecuadamente el funcionamiento de los sistemas y sus decisiones. La adhesión a principios y directrices éticas debe tomarse en serio y garantizar que los riesgos relacionados con los derechos humanos se mitiguen de manera suficiente antes de la distribución y despliegue de los sistemas de IA.

Controlados por las personas, que deben tener siempre el control y mantener la supervisión de los sistemas de IA, tanto durante su desarrollo como durante su uso. También creemos que cualquier decisión sobre cuestiones éticas que pueda afectar de manera fundamental a los derechos y a la dignidad de grupos y personas no debe externalizarse por completo a las máquinas. Para que las personas asuman plenamente el control es de enorme importancia que exista una sensibilización y una alfabetización digital adecuadas y generalizadas.

Uso adecuado. Debemos poner en marcha mecanismos que ayuden a evitar el uso excesivo de la tecnología y a depender demasiado de ella. Para conseguir una humanidad y unas vidas verdaderamente saludables es importante que preservemos y cuidemos de forma activa las cualidades y capacidades humanas esenciales para mantener nuestra creatividad, el dominio de las habilidades manuales y las capacidades cognitivas. Los sistemas de IA proporcionan un apoyo excelente a, pero

[2] La lista recoge la redacción elegida por el Grupo de Expertos de Alto Nivel de la UE sobre IA en sus Directrices Éticas de 2019 para una IA fiable (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai).
Existen dos excepciones: los términos *explicabilidad* y *responsabilidad* se añadieron a transparencia y rendición de cuentas, respectivamente. Hemos optado por hacerlo así para reflejar otras especificaciones que consideramos relevantes y que se incluyen en otras directrices oficiales.

su desarrollo y uso también tienen un coste. Debemos ser conscientes de los recursos medioambientales y humanos que se utilizan y se ven afectados. Los sistemas de IA no deben utilizarse más allá de lo necesario para alcanzar un objetivo legítimo. El uso excesivo de la tecnología puede acarrear costes medioambientales y sociales a largo plazo y no debe socavar nuestra confianza en el criterio humano.

Desarrollo y uso responsables. Los sistemas de IA deben desarrollarse y utilizarse con respeto a las leyes y normativas existentes, incluidas las relativas a los derechos humanos y los derechos de propiedad intelectual. Cuando existan lagunas normativas que creen vacíos legislativos debe tenerse en cuenta el espíritu de esas leyes. Como institución financiera que canaliza dinero hacia empresas locales, remotas y multinacionales, Triodos Bank cree que las empresas son las responsables últimas de desarrollar y utilizar sistemas de IA que cumplan los estándares de seguridad y solidez más altos y que defiendan valores éticos sólidos a través de la concienciación sobre nuestros derechos y deberes digitales individuales y colectivos. Aunque en la actualidad el desarrollo de los sistemas de IA más avanzados tiene lugar principalmente en el seno de entidades y corporaciones privadas, los avances en este ámbito revisten un interés colectivo mucho más amplio. Por eso debe establecerse una gobernanza adecuada de los sistemas de IA que implique a todas las partes interesadas pertinentes, para garantizar que los desarrollos en esta materia se llevan a cabo en beneficio del interés general.

3.2 Uso controvertido de la tecnología de IA

Algunas tecnologías y usos de la tecnología de IA se consideran muy problemáticos y plantean riesgos inaceptables. En algunos casos han sido prohibidos de manera expresa por las instituciones europeas a través de la Ley de Inteligencia Artificial de la UE. En Triodos Bank consideramos que los sistemas y usos de la IA siguientes no se ajustan a los principios de la IA responsable y deben ser firmemente condenados:

Armas autónomas letales. Las armas autónomas son sistemas de armamento que, una vez activados, pueden detectar y atacar de forma autónoma un objetivo sin la aprobación o intervención de un ser humano. Las armas autónomas letales no solo atentan contra el derecho a la vida, sino que se desarrollan y se utilizan sin marcos regulatorios estrictos. Vulneran el principio de supervisión humana en el uso de sistemas de IA y son un ejemplo censurable de externalización moral al delegar en las máquinas decisiones de vida o muerte[27].

Identificación biométrica en espacios públicos. La identificación biométrica, concretamente el reconocimiento facial en espacios de acceso público representa una vulneración sustancial del derecho a la intimidad. Las personas son identificadas sin su consentimiento y existe un riesgo alto para la seguridad de los datos, lo que puede llevar al fraude y a la usurpación de identidad. Además, esta tecnología puede utilizarse para la vigilancia masiva, lo que supone un riesgo alto para la libertad de expresión y asociación y para los derechos democráticos en general. El saber que somos observados u observadas puede cambiar nuestra forma de comportarnos y afectar a nuestra salud mental y a nuestro bienestar[28]. También se ha demostrado que su precisión varía en función del grupo demográfico, lo que crea riesgos muy elevados de discriminación y puede representar una amenaza para los derechos de la infancia[29].

Categorización biométrica y reconocimiento de emociones. Una tecnología y un suso muy problemáticos en su diseño y uso y en su proceso de recopilación de datos, que implica la extracción de datos biométricos de Internet —de redes sociales, por ejemplo— sin un consentimiento significativo. Categorizar a las personas en función de sus rasgos físicos presenta riesgos de discriminación muy elevados y a menudo se ha demostrado que carece de base científica. La detección de emociones también carece de una base científica sólida y fiable[30] y puedes ser enormemente perjudicial en contextos delicados, como en el laboral y en el educativo (según la normativa de la UE), así como en la aplicación de la ley, la justicia penal y el control de fronteras.

Puntuación social. Los sistemas de IA se utilizan también con fines de valoración o puntuación social para clasificar o evaluar a las personas en función de su comportamiento social o de características personales o de personalidad conocidas o esperadas[31]. Esos sistemas pueden dar lugar a un trato injusto y discriminatorio de personas y grupos de personas. También pueden comprometer la privacidad y conducir a la elaboración de perfiles basados en estereotipos y sesgos no reconocidos, con repercusiones importantes en los derechos democráticos y en el acceso a un nivel de vida adecuado. Por esas razones, estamos especialmente en contra de los sistemas predictivos y de elaboración de perfiles en ámbitos como la aplicación de la ley y la justicia penal, así como de las aprobaciones automáticas de créditos sin supervisión humana.

Manipulación cognitiva y del comportamiento. La tecnología impulsada por la IA puede desarrollarse o utilizarse para influir en el comportamiento humano —por ejemplo, las prácticas de marketing persuasivo y los sistemas diseñados para impulsar la adicción— y generan un riesgo alto de comportamientos adictivos. Otro tipo de práctica manipuladora es la difusión de información errónea, como los *deepfakes* (vídeos y audios falsos) y los medios sintéticos. Estas prácticas pueden tener consecuencias peligrosas y socavar, por ejemplo, los procesos democráticos y la confianza en las instituciones. Aunque es muy difícil

regular estas prácticas creemos que deberíamos abstenernos colectivamente de fomentar el desarrollo y el uso de sistemas de IA que puedan causar daños sustanciales a través de la manipulación cognitiva y del comportamiento.

Consideramos que las aplicaciones de IA mencionadas anteriormente pueden tener un impacto negativo grave en el bienestar de las personas, la salud de nuestro tejido social y el futuro de la humanidad en general y que las personas deben estar protegidas frente a ellas en todo el mundo. En el capítulo siguiente definiremos el papel que deben adoptar las instituciones financieras en la promoción de una IA responsable de acuerdo con los principios y consideraciones expuestos antes en relación con esas controvertidas aplicaciones.

La tecnología de IA y los temas de transición de Triodos Bank

La estrategia de impacto de Triodos Bank gira en torno a cinco temas de transición (alimentaria, energética, social, de recursos y del bienestar). La tecnología, y por tanto la tecnología de IA, puede desempeñar un papel vital en cada una de las transiciones. Sin embargo, queremos hacer hincapié en cómo afecta a las personas y a la sociedad en general.

Por eso la contemplamos desde la óptica de la transición del bienestar y de la transición social. Consideramos cómo afecta a nuestro tejido social y a nuestros cimientos como sociedad, y esto nos lleva a adoptar un enfoque preventivo. Abogamos por situar la dignidad humana en el centro del desarrollo tecnológico y crear las condiciones adecuadas para utilizarla de una forma saludable que favorezca el bienestar físico y mental. Colectivamente pasaríamos de una tecnología centrada en el ser humano a otra centrada en la humanidad. Esto significa que el diseño y el uso deben centrarse en reducir las desigualdades, en lugar de exacerbarlas, y en apoyar la construcción de sociedades justas, cohesionadas y pacíficas. Si no nos mantenemos firmes respecto a esos principios nos limitaremos a tener tecnología simplemente “porque hay que tenerla”.

4. El papel de las instituciones financieras

Las instituciones financieras son clave en nuestra economía por su doble papel de intermediación monetaria y su papel de ciudadanía corporativa y en ambos pueden desempeñar —y desempeñan— un papel fundamental en relación con las partes interesadas. Triodos Bank se toma muy en serio estos dos papeles y queremos destacar las maneras en que las instituciones financieras —con el bienestar de la sociedad como ADN de sus actividades— pueden fomentar prácticas responsables en relación con la tecnología de IA.

4.1 Las instituciones financieras como intermediarias monetarias. Expectativas sobre las empresas

Al canalizar el dinero hacia el sector privado, las instituciones financieras desempeñan un papel importante en el establecimiento de expectativas e incentivos para las empresas. Hasta la fecha, la tecnología de IA la desarrollan en gran medida empresas privadas y al mismo tiempo empresas de todo el mundo adoptan rápidamente sistemas de IA en sus actividades. La regulación de los sistemas de IA va con retraso y en vista de que su plena aplicación llevará tiempo, las empresas son, en gran medida, las que determinan las prácticas en torno al desarrollo y uso de los sistemas de IA.

Las instituciones financieras, como los bancos y las personas o entidades inversoras, suelen emplear dos fórmulas para fomentar las buenas prácticas empresariales. Pueden examinar a las empresas y seleccionar solo las que muestren prácticas suficientemente buenas para financiarlas e invertir en ellas o pueden comprometerse con las entidades en cartera y elevarlas hacia estándares más altos.
En el contexto de la tecnología de IA, creemos que las instituciones financieras responsables deben adoptar un enfoque preventivo y exigir que las empresas a las que financian e invierten demuestren su responsabilidad respecto a los riesgos de la tecnología que desarrollan y utilizan. Y para eso tendrán que avanzar en la dirección siguiente:

- **Compromiso con la IA responsable.** Las empresas que tienen una exposición sustancial a la tecnología de IA, como usuarias o como desarrolladoras, deberán declarar su compromiso público sobre la IA responsable. Es importante que estos documentos o declaraciones demuestren una concienciación considerable y que hagan referencia expresa a los impactos y riesgos adversos más relevantes para los derechos humanos de los sistemas que la empresa desarrolla o utiliza. También deben comprometerse con la ética en el diseño y con una tecnología que sitúe a la humanidad en el foco de sus actividades.

- **No llevar a cabo actividades perjudiciales.** Deben esperar y exigir que las empresas se comprometan a no utilizar, desarrollar o contribuir al desarrollo de sistemas de IA para usos muy controvertidos y que dispongan de sistemas adecuados de diligencia debida para garantizar que cumplen ese compromiso. Como mínimo, debe esperarse que las empresas cumplan la regulación existente sobre actividades controvertidas, pero las instituciones financieras tienen también la facultad de establecer estándares más estrictos que los exigidos por la ley. De ese modo lanzan un mensaje claro a las empresas. Un ejemplo claro de exigencia más allá de la regulación implicaría, por ejemplo, exigir la no participación en la producción y distribución de armas autónomas letales.

- **Luchar por la transparencia y la buena gobernanza.** Las instituciones financieras pueden exigir que las empresas sean transparentes respecto a los sistemas de IA que desarrollan o utilizan. Sin un nivel fundamental de transparencia es muy difícil llevar a cabo un escrutinio más exhaustivo. Las empresas que participan en la investigación y el desarrollo de sistemas de IA y proporcionan productos o servicios relacionados con la IA —en particular las que desarrollan los sistemas— deben ser capaces de explicar cómo se diseñaron y entrenaron, qué datos se utilizaron y cómo se probaron los modelos. Además, deben disponer de sólidos mecanismos de gobernanza para garantizar el desarrollo y uso responsables de los sistemas de IA y definir claramente la rendición de cuentas. Las empresas que tienen una implicación considerable en la tecnología de IA deben contar con un comité de ética específico que supervise el desarrollo y el uso responsables de los sistemas de IA en toda la empresa. El comité debe tener una responsabilidad y un poder de decisión claros y llevar a cabo las evaluaciones de impacto ético y de impacto sobre los derechos fundamentales recomendados.

- **Políticas específicas en vigor.** En función de la naturaleza del negocio, las instituciones financieras pueden exigir que las empresas que utilizan sistemas de IA dispongan de políticas específicas y mecanismos de gobernanza que cubran el uso de la tecnología de IA. Por ejemplo, las empresas que utilizan sistemas de IA con fines de marketing deberían hacerlo en el marco de una política sobre prácticas de marketing responsable que aborde de manera expresa los riesgos asociados al uso de sistemas de IA.

Comprometerse con las empresas para mejorar sus actividades y abogar por las mejores prácticas en el sector debería ser el paso siguiente, tanto para las entidades de inversión como para los bancos que conceden préstamos a empresas, que pueden entablar un diálogo constructivo sobre la necesidad de impartir una formación adecuada en materia de alfabetización digital y, en particular, sobre el uso de la tecnología de IA y los riesgos relativos a los las personas que trabajan en los distintos niveles de la empresa.

También podrían debatir la necesidad de proporcionar directrices claras para las personas usuarias de aplicaciones basadas en IA. Si la IA se utiliza para reducir costes y aumentar la eficiencia a expensas de los puestos de trabajo, las empresas deben explicar con transparencia cómo piensan aplicar las salvaguardas pertinentes para las personas despedidas. Además, las y los inversores podrían animar a las empresas en las que invierten a adoptar una postura clara sobre el rápido desarrollo de la tecnología y apoyar iniciativas que exijan normativas y la interrupción del desarrollo de las tecnologías más avanzadas hasta que se establezcan regulaciones exhaustivas y una gobernanza adecuada. Por supuesto, los requisitos para las empresas implicadas en el desarrollo o uso de la tecnología de IA deben ser razonables y proporcionales al tamaño de la empresa.

Qué espera Triodos Bank de las empresas

Triodos Bank es un banco con valores y un inversor de impacto responsable que aplica unos estándares éticos altos y unas expectativas elevadas a las empresas a las que financia y en las que invierte. Creemos que ha llegado el momento de hacer lo mismo con las empresas que desarrollan o utilizan tecnología de IA.

Triodos Bank espera que, como mínimo, las empresas a las que financia y en las que invierte cumplan los requisitos de la Ley de Inteligencia Artificial de la UE una vez se aplique ese reglamento.
Los Estándares Mínimos que Triodos Bank aplica ya a las empresas que financia y en las que invierte incluyen criterios de exclusión que cubren implícitamente la mayoría de los usos controvertidos de la tecnología de IA. Eso supone, por ejemplo, excluir de su universo de inversión a las empresas que faciliten el desarrollo y el uso de sistemas de IA en armas autónomas (estándares mínimos sobre armas, armamento y municiones), la puntuación social, el reconocimiento facial y de emociones en entornos sensibles, o la identificación biométrica en tiempo real y a distancia en espacios públicos que pueda conducir a la vigilancia masiva (estándares mínimos sobre derechos humanos). La próxima actualización de los Estándares Mínimos hará referencia expresa a esas prácticas.

Analizamos a las empresas para asegurarnos de que cumplen nuestros estándares mínimos según nuestro saber y entender. Si identificamos que alguna empresa de nuestra cartera está implicada en controversias, ponemos en marcha un proceso para aclarar la situación que puede llevar a su exclusión de la cartera. Las controversias relacionadas con la IA pueden incluir el suministro de los sistemas para usos muy controvertidos, la exportación de tecnología crítica a estados o entidades sancionados, el sesgo algorítmico y la privacidad y seguridad de los datos, así como las infracciones de los derechos de propiedad intelectual y las normativas y prácticas manipuladoras impulsadas por la IA. Aunque algunos de nuestros estándares van más allá de lo que exige la normativa, sobre todo en el caso de las armas autónomas (que no están cubiertas por la Ley de Inteligencia Artificial de la UE), por el momento no esperamos formalmente que las empresas cumplan requisitos más allá de los establecidos por esa ley y otras normativas.

Sin embargo, a medida que mejoren nuestro propio conocimiento y nuestras prácticas internas en la materia, tenemos la intención de examinar más de cerca las políticas y programas relacionados con la IA de las empresas. Nuestro objetivo es animarlas a que adopten las prácticas descritas y por eso Triodos Investment Management, la división dedicada a la gestión de inversiones de Triodos Bank, se ha incorporado recientemente a la Coalición de Impacto Colectivo para una IA Ética de la World Benchmarking Alliance. Esperamos aprovechar esta interacción colaborativa para impulsar las prácticas responsables de la IA de acuerdo con los principios y recomendaciones descritos en este documento.
-

4.2 Las instituciones financieras y su ciudadanía corporativa. Actividades propias y voz pública

Las instituciones financieras, como todas las empresas, son ciudadanas corporativas. Tienen el deber de servir a sus clientes, pero también una responsabilidad más amplia ante la sociedad. Además, pueden dar a conocer su opinión de forma abierta y transparente en el ámbito público y ante las instituciones.

Cuando se trata de las actividades propias de la institución financiera, los bancos y entidades de inversión no deben subestimar su papel para conseguir que los sistemas de IA se desarrollen y se utilicen de acuerdo con los estándares éticos y técnicos más estrictos. El sector financiero ha adoptado la IA cada vez en mayor medida para varios usos y aplicaciones. Cuando se aplican a los procesos de concesión de créditos o a la puntuación crediticia, corren un riesgo alto de generar resultados sesgados con la perpetuación de prejuicios y la generación de discriminaciones en el acceso al crédito y en los precios[32, 33].

El uso de esos datos por parte de las entidades financieras genera preocupaciones importantes. Según el Banco Central de los Países Bajos cuando las instituciones financieras utilizan los datos de los clientes como activo comercial pueden socavar la confianza en las instituciones y en el sistema financiero[34], incluso si se hace dentro de los marcos legales.

La ciberseguridad es otro motivo de preocupación. La IA Generativa puede generar mensajes de phishing más sofisticados y facilitar a agentes maliciosos la suplantación de identidad de las personas. Además, la negociación bursátil basada en algoritmos (*algorithmic trading*) ya representa una fuente importante de posible inestabilidad para el sistema financiero, que podría agravarse aún más con modelos más potentes. Lo mismo sucede con los sistemas de IA aplicados a la gestión de riesgos[35].

Por lo tanto, las instituciones financieras deberían estar a la altura de los estándares que exigen a sus empresas participadas y financiadas y esforzarse por conseguir una gobernanza adecuada en torno a la tecnología de IA, tanto si solo la utilizan como si también la desarrollan.

Sistemas de IA en Triodos Bank

Triodos Bank utiliza sistemas de IA en varios departamentos para la supervisión de transacciones, la lucha contra el blanqueo de capitales y la detección del fraude, así como para la verificación de la identidad mediante el reconocimiento facial como parte de la incorporación digital de clientes, para la recopilación de información en la investigación de empresas y para la traducción inicial de documentos de marketing.

Nuestras empresas participadas y clientes empresariales pueden utilizar sistemas de IA en una gama amplia de ámbitos para acelerar su impacto positivo, como la gestión inteligente de la energía, la supervisión medioambiental, el diseño eficiente de los recursos, el seguimiento y la optimización de la cadena de suministro, el diagnóstico y la investigación de fármacos. Triodos Bank ha formulado sus propios principios internos para una IA ética junto con directrices para las personas trabajadoras, con el fin de mejorar la alfabetización digital y concienciar sobre las oportunidades y los riesgos relacionados con el uso de los sistemas de IA.

Desde nuestra ciudadanía corporativa, y con una influencia potente e innegable, las instituciones financieras podemos expresar opiniones sobre los riesgos financieros y sociales más amplios en relación con el desarrollo y la difusión de la tecnología de IA. Pueden aprovechar muchas oportunidades para influir en los organismos reguladores y en quienes toman decisiones, siempre que sean transparentes y coherentes al exponer sus puntos de vista, tanto en privado como en público. En el cuadro de texto siguiente se incluye un ejemplo de la posición de Triodos Bank respecto a la Ley de Inteligencia Artificial aprobada recientemente por las instituciones de la UE.

Postura de Triodos Bank respecto a la Ley de Inteligencia Artificial de la UE

Triodos Bank opera directamente en varios países de la UE y en el Reino Unido y desarrolla su actividad en todo el mundo a través de sus operaciones de inversión. Como tal, el banco tiene interés en los desarrollos normativos relacionados con la tecnología de IA. La Ley de Inteligencia Artificial de la UE ha recorrido un camino largo en la regulación de las tecnologías impulsadas por IA, con la introducción de normas para mitigar los riesgos relacionados con esos sistemas y con la preparación del terreno para crear autoridades de supervisión específicas.

Aunque Triodos Bank no tiene previsto por el momento adoptar un papel activo en la defensa de cuestiones relacionados con la IA, somos conscientes de los riesgos y problemas relacionados con los sistemas y que no se abordan plenamente en la normativa. En particular, destacan algunos puntos:

- Triodos Bank no está de acuerdo con la exención regulatoria de los sistemas de IA en materia de seguridad nacional por que deja margen para el desarrollo y uso de armas autónomas impulsadas por IA.
- Triodos Bank cree que los mismos requisitos que se aplican a los sistemas de IA utilizados en la UE deben aplicarse a los sistemas de IA desarrollados en la UE y exportados fuera de sus fronteras para garantizar que las empresas y gobiernos de la UE no se beneficien de exportar tecnología potencialmente abusiva.
- Triodos Bank no aprueba el uso de la identificación biométrica a distancia en espacios públicos ni el reconocimiento de emociones en entornos sensibles por fuerzas del orden o para juicios penales y control de fronteras.
- Triodos Bank espera que los próximos pasos en la aplicación de esta regulación conduzcan a una definición clara de lo que constituye un sistema de IA de alto riesgo, al abordar las lagunas normativas que actualmente permiten a las empresas decidir si esta normativa es o no aplicable a los sistemas de IA que desarrollan.

4. Llamamiento a la acción

El uso de los sistemas de IA ya se ha generalizado y tienen muchas aplicaciones valiosas. La tecnología de IA no es intrínsecamente buena ni mala, pero el uso generalizado de los sistemas de IA es una fuerza transformadora que ya afecta a nuestras economías y que debe regularse de manera consciente. Aunque se introducen cada vez más normativas al respecto, las empresas todavía asumen en gran medida la responsabilidad de gestionar los riesgos e impactos que en materia de derechos humanos y de ética empresarial generan en sus grupos de interés a la vez que desarrollan y despliegan sistemas de IA centrados en la humanidad.

Triodos Bank ha desarrollado sus propios puntos de vista y sus expectativas respecto a las empresas a las que financia y en las que invierte, con los que pretende garantizar el compromiso con la IA responsable y las buenas prácticas en su desarrollo y su uso. Triodos Bank también trabaja en la mejora de sus mecanismos internos y en el conocimiento de la tecnología de IA para garantizar un uso responsable.

Reconocemos el papel que las instituciones financieras pueden desempeñar para establecer un desarrollo y uso responsables de la tecnología de IA. Triodos Bank hace un llamamiento a otras instituciones financieras para que se abstengan de incentivar o aprovecharse ciegamente de un desarrollo rápido e irresponsable de esta tecnología en el contexto de una industria no regulada y altamente competitiva. En su lugar, deben adoptar un enfoque preventivo, con expectativas claras y el compromiso con una tecnología de IA responsable. Es responsabilidad común y colectiva comprender el poder de la tecnología de la IA y sus consecuencias para la humanidad.

Referencias

[1] Gugerty, L. (2006). *Newell and Simon's logic theorist: historical background and impact on cognitive modeling*. Actas de la reunión anual de la Human Factors and Ergonomics Society, Vol. 50, Nº. 9, pp. 880-884). Sage CA: Los Angeles, CA: SAGE Publications.

[2] Medium (4 de diciembre de 2020). *The first of its kind AI Model- Samuel's Checkers Playing Program*.

[3] OCDE (2024), Recomendación del Consejo sobre la Inteligencia Artificial, OCDE/LEGAL/0449.

[4] Washington Post (26 de marzo de 2023). *There's no such thing as artificial intelligence*.

[5] Harvard Business Review (2021). *AI regulation is coming*.

[6] ONU (26 de octubre de 2023). *New UN Advisory Body aims to harness AI for the common good*.

[7] Politico (31 de octubre de 2023). *UK, US slated to announce AI safety partnership*.

[8] Comisión Europea (30 de octubre de 2023). *Hiroshima Process International Code of Conduct for Organisations Developing Advanced AI Systems*.

[9] BBC News (10 de octubre de 2023). *Warning AI industry could use as much energy as the Netherlands*.

[10] S&P Global Commodity Insights (16 de octubre de 2023). *Power of AI: Wild predictions of power demand from AI put industry on edge*.

[11] Li, P. et al. (2023). *Making AI less" thirsty": Uncovering and addressing the secret water footprint of AI models*. arXiv preprint arXiv:2304.03271.

[12] Forti, V. et al. (2020). *The global e-waste monitor 2020*. United Nations University (UNU), International Telecommunication Union (ITU) & International Solid Waste Association (ISWA), Bonn/Ginebra/Rotterdam, 120.

[13] Naciones Unidas, Asamblea General (1948). *Declaración Universal de Derechos Humanos*.

[14] Unión Europea, Convención Europea (2000). *Carta de los Derechos Fundamentales de la Unión Europea*.

[15] Naciones Unidas, Asamblea General (1966). *Pacto Internacional de Derechos Civiles y Políticos*.

[16] Varona, D., Suárez, J.L. (2022). *Discrimination, bias, fairness, and trustworthy AI*. Applied Sciences, 12(12):5826.

[17] Harvard T.H. Chan School of Public Health (March 12, 2021). *Algorithmic bias in health care exacerbates social inequities — How to prevent it*.

[18] MIT Technology Review (17 de junio de 2021). *Bias isn't the only problem with credit scores—and no, AI can't help*.

[19] Dresp-Langley, B. (2023). *The weaponization of artificial intelligence: What the public needs to be aware of*. Frontiers in Artificial Intelligence, 6:1154184.

[20] Forbes (6 de enero de 2021). *The U.S. army's robot tanks will make great bait*.

[21] Treleaven, P. et al. (2023). *The future of cybercrime: AI and emerging technologies are creating a cybercrime tsunami*. Social Science Research Network (SSRN).

[22] OSCE (2020). *Global Conference for Media Freedom: Freedom of the media and artificial intelligence*.

[23] Harvard Business Review (7 de abril de 2023). *Generative AI has an intellectual property problem*.

[24] Financial Times (6 de noviembre de 2023). *What AI means for ESG*.

[25] Noema (13 de octubre de 2022). *The exploited labor behind artificial intelligence*.

[26] Council on Foreign Relations (25 de octubre de 2023). *Governing artificial intelligence: A conversation with Rumman Chowdhury*.

[27] Pax (17 de octubre de 2023). *Increasing complexity - Legal and moral implications of trends in autonomy in weapons systems*.

[28] Smith, M. J. et al. (1992). *Employee stress and health complaints in jobs with and without electronic performance monitoring*. Applied Ergonomics, 23(1), 17-27.

[29] UNICEF (2019), *Faces, fingerprints and feet*.

[30] Barrett, L. F. et al. (2019). *Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements*. Psychological Science in the Public Interest, 20(1), 1-68.

[31] Comisión Europea (2021), *Propuesta de reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión*.

[32] Garcia, A. C. B. et al. (2023). *Algorithmic discrimination in the credit domain: What do we know about it?*. AI & Society, 1-40.

[33] MIT Technology Review (17 de junio de 2021). *Bias isn't the only problem with credit scores—and no, AI can't help*.

[34] De Nederlandsche Bank (2019). *General principles for the use of artificial intelligence in the financial sector*.

[35] Danielsson, J. et al. (2022). *Artificial intelligence and systemic risk*. Journal of Banking & Finance, 140, 106290.

Address
Hoofdstraat 10, Driebergen-Rijsenburg
PO Box 55
3700 AB Zeist, The Netherlands
Telephone +31 (0)30 693 65 00
www.triodos.com
www.triodos-im.com
www.triodos.com/regenerative-money-centre

Published
July 2024

Text
Federica Masut and Johanna Schmidt, Triodos Bank

Design and layout
PI&Q, Zeist

Acknowledgments
Triodos Bank would like to thank Iris Muis (Data School, Utrecht University) for the insightful feedback on this paper.

Triodos Bank