GRUPO INDEPENDIENTE DE EXPERTOS DE ALTO NIVEL SOBRE INTELIGENCIA ARTIFICIAL

CREADO POR LA COMISIÓN EUROPEA EN JUNIO DE 2018

DIRECTRICES ÉTICAS PARA UNA IA FIABLE

DIRECTRICES ÉTICAS para una IA FIABLE

Grupo de expertos de alto nivel sobre inteligencia artificial

Este documento ha sido redactado por el Grupo de expertos de alto nivel sobre inteligencia artificial (IA). Los miembros del Grupo de expertos citados en este documento respaldan el marco general para una IA fiable descrito en las presentes directrices, aunque no están necesariamente de acuerdo con todas y cada una de las afirmaciones que se realizan en ellas..

La lista para la evaluación de la fiabilidad de la IA que se expone en el capítulo III de este documento se someterá a una fase de experimentación con carácter piloto por las partes interesadas, a fin de recabar sus opiniones prácticas. A principios de 2020 se presentará a la Comisión Europea una versión revisada de dicha lista de evaluación, teniendo en cuenta los comentarios recogidos durante la fase de experimentación.

El Grupo de expertos de alto nivel sobre IA es un grupo de expertos independientes constituido por la Comisión Europea en junio de 2018.

Contacto Nathalie Smuha - Coordinadora del Grupo de expertos de alto nivel sobre IA
Correo electrónico CNECT-HLG-AI@ec.europa.eu

Comisión Europea
B - 1049 BRUSELAS

Documento publicado el X de abril de 2019.

El 18 de diciembre de 2018 se publicó un primer borrador de este documento, que se sometió a un proceso de consulta abierta a través del cual se recogieron comentarios de más de 500 participantes. Los autores desean expresar su sincero agradecimiento a todas las personas que realizaron aportaciones al primer borrador del documento; dichas aportaciones se tuvieron en cuenta durante la elaboración de esta versión revisada.

ÍNDICE

RESUMEN **2**

A. INTRODUCCIÓN **5**

B. UN MARCO PARA UNA IA FIABLE **7**

I. Capítulo I: Fundamentos de una IA fiable **11**
1. Los derechos fundamentales como derechos morales y legales 11
2. De los derechos fundamentales a los principios éticos 12

II. Capítulo II: Realización de la IA fiable **17**
1. Requisitos de una IA fiable 17
2. Métodos técnicos y no técnicos para hacer realidad la IA fiable 25

III. Capítulo III: Evaluación de la IA fiable **30**

C. EJEMPLOS DE OPORTUNIDADES Y PREOCUPACIONES FUNDAMENTALES QUE PLANTEA LA IA 42

D. CONCLUSIÓN **46**

GLOSARIO **48**

RESUMEN

1) El objetivo de las presentes directrices es promover una inteligencia artificial fiable. La fiabilidad de la inteligencia artificial (IA) se apoya en **tres componentes** que deben satisfacerse a lo largo de todo el ciclo de vida del sistema: a) la IA debe ser **lícita**, es decir, cumplir todas las leyes y reglamentos aplicables; b) ha de ser **ética**, de modo que se garantice el respeto de los principios y valores éticos; y c) debe ser **robusta**, tanto desde el punto de vista técnico como social, puesto que los sistemas de IA, incluso si las intenciones son buenas, pueden provocar daños accidentales. Cada uno de estos componentes es en sí mismo necesario pero no suficiente para el logro de una IA fiable. Lo ideal es que todos ellos actúen en armonía y de manera simultánea. En el caso de que surjan tensiones entre ellos en la práctica, la sociedad deberá esforzarse por resolverlas.

2) Estas directrices establecen un **marco para conseguir una IA fiable.** Dicho marco no aborda explícitamente el primero de los tres componentes expuestos de la inteligencia artificial (IA lícita)[1]. En lugar de ello, pretende ofrecer orientaciones sobre el fomento y la garantía de una IA ética y robusta (los componentes segundo y tercero). Las directrices, que van dirigidas a todas las partes interesadas, buscan ofrecer algo más que una simple lista de principios éticos; para ello, proporcionan orientación sobre cómo poner en práctica esos principios en los sistemas sociotécnicos. Las orientaciones se ofrecen en tres niveles de abstracción, desde el capítulo I (el más abstracto) al III (el más concreto). Concluyen con ejemplos de las oportunidades y preocupaciones fundamentales que plantean los sistemas de IA.

   I. Partiendo de un enfoque basado en los derechos fundamentales, en el capítulo I se identifican los **principios éticos** y sus valores conexos que deben respetarse en el desarrollo, despliegue y utilización de los sistemas de IA.

   Orientaciones clave derivadas del capítulo I:

   - ✓ Desarrollar, desplegar y utilizar los sistemas de IA respetando los principios éticos de: *respeto de la autonomía humana, prevención del daño, equidad y explicabilidad.* Reconocer y abordar las tensiones que pueden surgir entre estos principios.
   - ✓ Prestar una atención especial a las situaciones que afecten a los grupos más vulnerables, como los niños, las personas con discapacidad y otras que se hayan visto históricamente desfavorecidas o que se encuentren en riesgo de exclusión, así como a las situaciones caracterizadas por asimetrías de poder o de información, como las que pueden producirse entre empresarios y trabajadores o entre empresas y consumidores[2].
   - ✓ Reconocer y tener presente que, pese a que aportan beneficios sustanciales a las personas y a la sociedad, los sistemas de IA también entrañan determinados riesgos y pueden tener efectos negativos, algunos de los cuales pueden resultar difíciles de prever, identificar o medir (por ejemplo, sobre la democracia, el estado de Derecho y la justicia distributiva, o sobre la propia mente humana). Adoptar medidas adecuadas para mitigar estos riesgos cuando proceda; dichas medidas deberán ser proporcionales a la magnitud del riesgo.

   II. A partir de lo expuesto en el capítulo I, el capítulo II ofrece orientaciones sobre cómo lograr una IA fiable, enumerando **siete requisitos** que deben cumplir los sistemas de IA para ello. Para aplicar estas orientaciones se pueden utilizar tanto métodos técnicos como de otro tipo.

---

[1] Todas las declaraciones normativas recogidas en este documento tienen la finalidad de brindar orientaciones de cara al logro de los componentes segundo y tercero de una IA fiable (IA ética y robusta). Por lo tanto, dichas declaraciones no pretenden ofrecer asesoramiento jurídico ni orientaciones sobre el cumplimiento de las leyes aplicables, aunque se reconoce que muchas de ellas figuran recogidas en cierta medida en las leyes existentes. Véanse, en ese sentido, los puntos 21 y ss.

[2] Véanse los artículos 24 a 27 de la Carta de los Derechos Fundamentales de la Unión Europea (Carta de la UE), que tratan sobre los derechos de los niños y de las personas mayores, la integración de las personas con discapacidad y los derechos de los trabajadores. Véase también el artículo 38 relativo a la protección de los consumidores.

Orientaciones clave derivadas del capítulo II:

- ✓ Garantizar que el desarrollo, despliegue y utilización de los sistemas de IA cumpla los requisitos para una IA fiable: 1) acción y supervisión humanas, 2) solidez técnica y seguridad, 3) gestión de la privacidad y de los datos, 4) transparencia, 5) diversidad, no discriminación y equidad, 6) bienestar ambiental y social, y 7) rendición de cuentas.
- ✓ Para garantizar el cumplimiento de estos requisitos, se deberá estudiar la posibilidad de emplear tanto métodos técnicos como no técnicos.
- ✓ Impulsar la investigación y la innovación para ayudar a evaluar los sistemas de IA y a promover el cumplimiento de los requisitos; divulgar los resultados y las preguntas de interpretación abierta al público en general, y formar sistemáticamente a una nueva generación de especialistas en ética de la IA.
- ✓ Comunicar información a las partes interesadas, de un modo claro y proactivo, sobre las capacidades y limitaciones de los sistemas de IA, posibilitando el establecimiento de expectativas realistas, así como sobre el modo en que se cumplen los requisitos. Ser transparentes acerca del hecho de que se está trabajando con un sistema de IA.
- ✓ Facilitar la trazabilidad y la auditabilidad de los sistemas de IA, especialmente en contextos o situaciones críticos.
- ✓ Implicar a las partes interesadas en todo el ciclo de vida de los sistemas de IA. Promover la formación y la educación, de manera que todas las partes interesadas sean conocedoras de la IA fiable y reciban formación en la materia.
- ✓ Ser conscientes de que pueden existir tensiones fundamentales entre los diferentes principios y requisitos. Identificar, evaluar, documentar y comunicar constantemente este tipo de tensiones y sus soluciones.

III. El capítulo III ofrece una lista concreta y no exhaustiva para la evaluación de la fiabilidad de la IA, con el objetivo de poner en práctica los requisitos descritos en el capítulo II. Dicha **lista de evaluación** deberá adaptarse al caso específico de utilización del sistema de IA[3].

Orientaciones clave derivadas del capítulo III:

- ✓ Adoptar una evaluación de la fiabilidad de la IA al desarrollar, desplegar o utilizar sistemas de IA, y adaptarla al caso de uso específico en el que se aplique dicho sistema.
- ✓ Tener presente que este tipo de listas de evaluación nunca pueden ser exhaustivas. Garantizar la fiabilidad de la IA no consiste en marcar casillas de verificación, sino en identificar y aplicar constantemente requisitos, evaluar soluciones y asegurar mejores resultados a lo largo de todo el ciclo de vida del sistema de IA, implicando a las partes interesadas en el proceso.

3) La sección final del documento tiene por objetivo concretar algunas de las cuestiones abordadas a lo largo del marco expuesto, ofreciendo ejemplos de oportunidades beneficiosas que se deberían perseguir y de preocupaciones cruciales que plantean los sistemas de IA y que deberían ser objeto de un estudio pormenorizado.

4) Pese a que estas directrices pretenden ofrecer orientaciones sobre las aplicaciones de la IA en general y establecer una base horizontal para lograr una IA fiable, las diferentes situaciones plantean desafíos distintos. Por lo tanto, se debería explorar si, además de este marco horizontal, puede ser necesario un enfoque sectorial dado que la aplicación de los sistemas de AI depende del contexto.

[3] En consonancia con el ámbito de aplicación del marco expuesto en el punto 2, esta lista de evaluación no proporciona consejo alguno sobre cómo garantizar el cumplimiento de la normativa (IA lícita), sino que se limita a ofrecer orientaciones de cara al cumplimiento de los componentes segundo y tercero de la IA fiable (IA ética y robusta).

5) Estas directrices no aspiran a reemplazar ninguna política o reglamento actual o futuro, ni a impedir su introducción. Deben considerarse un documento vivo, que habrá de revisarse y actualizarse a lo largo del tiempo para garantizar que sigan siendo pertinentes a medida que evolucionen la tecnología, el entorno social y nuestros propios conocimientos. Este documento ha sido concebido como un punto de partida para el debate sobre «Una IA fiable para Europa»[4]. Más allá de Europa, las directrices también buscan estimular la investigación, la reflexión y el debate sobre un marco ético para los sistemas de IA a escala mundial.

[4] Se pretende aplicar este ideal a los sistemas de IA desarrollados, desplegados y utilizados en los Estados miembros de la UE, así como a los sistemas desarrollados o producidos en otros lugares pero que se desplieguen y utilicen en la UE. Las referencias a «Europa» en este documento se entenderán hechas a los Estados miembros de la UE. No obstante, la aspiración es que estas directrices también resulten pertinentes fuera de la Unión. En este sentido, cabe señalar que tanto Noruega como Suiza participan en el plan coordinado sobre la inteligencia artificial aprobado y publicado en diciembre de 2018 por la Comisión y los Estados miembros.

A. INTRODUCCIÓN

6) En sus Comunicaciones de 25 de abril de 2018 y 7 de diciembre de 2018, la Comisión Europea (la Comisión) definió su visión acerca de la inteligencia artificial (IA), una visión que apoya una IA ética, segura y vanguardista «made in Europe»[5]. La visión de la Comisión se apoya en tres pilares: i) aumentar las inversiones públicas y privadas en IA para impulsar su adopción, ii) prepararse para los cambios socioeconómicos y iii) garantizar un marco ético y legal adecuado para fortalecer los valores europeos.

7) Para contribuir a hacer realidad esta visión, la Comisión creó el Grupo de expertos de alto nivel sobre inteligencia artificial, un grupo independiente al que se encomendó la redacción de dos documentos: 1) Directrices éticas sobre IA, y 2) Recomendaciones sobre política e inversión.

8) Este documento recoge las directrices éticas sobre la IA, revisadas tras las deliberaciones de nuestro Grupo en vista de las observaciones recibidas a través del proceso de consulta pública sobre el borrador divulgado el 18 de diciembre de 2018. Además, se apoya en el trabajo del Grupo europeo de ética de la ciencia y de las nuevas tecnologías[6] y se inspira en otras iniciativas similares[7].

9) A lo largo de los últimos meses, los 52 miembros de nuestro Grupo nos hemos reunido, debatido e interactuado desde el compromiso con el lema europeo: «Unida en la diversidad». Creemos que la IA tiene el potencial de transformar significativamente la sociedad. La IA no es un fin en sí mismo, sino un medio prometedor para favorecer la prosperidad humana y, de ese modo, mejorar el bienestar individual y social y el bien común, además de traer consigo progreso e innovación. En particular, los sistemas de IA pueden ayudar a facilitar el logro de los Objetivos de Desarrollo Sostenible de las Naciones Unidas, como la promoción del equilibrio entre mujeres y hombres y la lucha contra el cambio climático, la racionalización del uso que los seres humanos hacemos de los recursos naturales, la mejora de la salud, la movilidad y los procesos de producción y el seguimiento de los avances en los indicadores de sostenibilidad y cohesión social.

10) Para ello, es necesario que los sistemas de IA[8] se **centren en las personas** y se fundamenten en el compromiso de utilizarlos al servicio de la humanidad y del bien común, con el objetivo de mejorar el bienestar y la libertad de los seres humanos. Pese a que ofrecen magníficas oportunidades, los sistemas de IA también conllevan determinados riesgos a los que es preciso hacer frente de manera adecuada y proporcionada. Ahora tenemos una oportunidad muy importante para influir en su desarrollo. Queremos asegurarnos de poder confiar en los entornos sociotécnicos en los que se integran dichos sistemas, y deseamos que los productores de los sistemas de IA obtengan una ventaja competitiva mediante la incorporación de una IA fiable en sus productos y servicios. Esto implica tratar de **maximizar los beneficios de los sistemas de IA** y, al mismo tiempo, **prevenir y minimizar sus riesgos**.

11) En un contexto de rápido cambio tecnológico, creemos que es esencial que la confianza siga siendo el principal cimiento en el que se asienten las sociedades, comunidades y economías, así como el desarrollo sostenible. Por lo tanto, identificamos que la **IA fiable constituye nuestra aspiración fundacional,** puesto que los seres humanos y las comunidades solamente podrán confiar en el desarrollo tecnológico y en sus aplicaciones si contamos con un marco claro y detallado para garantizar su fiabilidad.

---

5 COM(2018)237 y COM(2018)795. Obsérvese que la expresión «made in Europe» se utiliza a lo largo de toda la Comunicación de la Comisión. Sin embargo, el ámbito de aplicación de estas directrices no pretende abarcar únicamente los sistemas de IA desarrollados en Europa, sino también los que sean desarrollados en otros lugares pero se desplieguen o utilicen en Europa. Por lo tanto, a lo largo de este documento nos referiremos a la promoción de una IA fiable «para» Europa.

6 El Grupo europeo de ética de la ciencia y de las nuevas tecnologías es un grupo consultivo de la Comisión.

7 Véase la sección 3.3 del documento COM(2018)237.

8 El glosario que figura al final de este documento ofrece una definición de los sistemas de IA a efectos del presente documento. Esta definición se desarrolla con mayor profundidad en un documento específico elaborado por el Grupo de expertos de alto nivel sobre inteligencia artificial que acompaña a estas directrices,, titulado «Una definición de la inteligencia artificial: Principales capacidades y disciplinas científicas».

12) Este es el camino que, a nuestro entender, debería seguir Europa para posicionarse como centro de desarrollo y como líder en el campo de una tecnología ética y de vanguardia. A través de una IA fiable, los ciudadanos europeos podremos tratar de cosechar los beneficios de esa tecnología de un modo acorde con nuestros valores fundacionales: respeto de los derechos humanos, democracia y estado de Derecho.

Una IA fiable

13) La fiabilidad es un requisito previo para que las personas y sociedades desarrollen, desplieguen y utilicen sistemas de IA. Si estos sistemas —y las personas que se encuentran detrás de ellos— no demuestran ser merecedores de confianza, pueden producirse consecuencias no deseadas que obstaculicen su adopción, impidiendo el logro de los enormes beneficios económicos y sociales que pueden acarrear los sistemas de IA. Para que Europa pueda obtener esos beneficios, nuestra visión consiste en establecer la ética como pilar fundamental para garantizar y expandir la IA fiable.

14) La confianza en el desarrollo, despliegue y utilización de sistemas de IA no concierne únicamente a las propiedades inherentes a esta tecnología, sino también a las cualidades de los sistemas sociotécnicos en los que se aplica la IA[9]. De forma análoga a las cuestiones relativas a la (pérdida de) confianza en la aviación, la energía nuclear o la seguridad alimentaria, no son simplemente los componentes del sistema de IA los que pueden generar (o no) confianza, sino el sistema en su contexto global. Por lo tanto, los esfuerzos dirigidos a garantizar la fiabilidad de la IA no solo atañen a la confianza que suscita el propio sistema de IA, sino que requieren un enfoque integral y sistémico que abarque la fiabilidad de todos los agentes y procesos que forman parte del contexto sociotécnico en el que se enmarca el sistema a lo largo de todo su ciclo de vida.

15) La fiabilidad de la inteligencia artificial (IA) se apoya en **tres componentes** que deben satisfacerse a lo largo de todo el ciclo de vida del sistema:

1. la IA debe ser **lícita**, de modo que se garantice el respeto de todas las leyes y reglamentos aplicables;
2. también ha de ser **ética**, es decir, asegurar el cumplimiento de los principios y valores éticos; y, finalmente,
3. debe ser **robusta**, tanto desde el punto de vista técnico como social, puesto que los sistemas de IA, incluso si las intenciones son buenas, pueden provocar daños accidentales.

16) Cada uno de estos tres componentes es necesario pero no suficiente para lograr una IA fiable[10]. Lo ideal es que todos ellos actúen en armonía y de manera simultánea. En la práctica, sin embargo, pueden surgir tensiones entre estos elementos (por ejemplo, en ocasiones el ámbito de aplicación y el contenido de la legislación existente pueden no ser coherentes con las normas éticas). Tenemos la responsabilidad individual y colectiva como sociedad de trabajar para garantizar que los tres componentes contribuyan a garantizar la fiabilidad de la IA[11].

17) Un enfoque basado en la fiabilidad es clave para una «competitividad responsable», dado que permite sentar las bases para que todos los afectados por los sistemas de IA puedan confiar en que su diseño, su desarrollo y su utilización son lícitos, éticos y robustos. Las presentes directrices pretenden fomentar la innovación responsable y sostenible en el campo de la IA en Europa. Su finalidad es convertir la ética en un pilar fundamental para desarrollar un enfoque único con respecto a la IA, que busque beneficiar, empoderar y proteger tanto la prosperidad humana a nivel individual como el bien común de la sociedad. A nuestro juicio, esto permitirá que Europa se posicione como líder mundial de una IA de vanguardia, merecedora de nuestra confianza individual y colectiva. Será imprescindible garantizar dicha confianza para que los ciudadanos

9 Estos sistemas abarcan personas, agentes estatales, corporaciones, infraestructura, programas informáticos, protocolos, normas, gobernanza, leyes existentes, mecanismos de supervisión, estructuras de incentivos, procedimientos de auditoría, informes sobre buenas prácticas, etc.

10 Esto no excluye el hecho de que pueden surgir condiciones necesarias adicionales.

11 Esto significa asimismo que el poder legislativo o los responsables de la formulación de políticas pueden necesitar revisar la idoneidad de la legislación existente cuando esta no esté en consonancia con los principios éticos.

europeos puedan sacar el máximo provecho de los sistemas de IA, sabiendo al mismo tiempo que se han adoptado las medidas necesarias para protegerlos frente a los posibles riesgos que conlleva esta tecnología.

18) El uso de sistemas de IA no se detiene en las fronteras nacionales, y sus efectos tampoco. En consecuencia, es preciso adoptar soluciones globales para las oportunidades y desafíos globales que trae consigo la IA. Por lo tanto, alentamos a todas las partes interesadas a trabajar en pos de la creación de un marco global para una IA fiable, alcanzar un consenso internacional y, al mismo tiempo, promover y defender nuestro enfoque fundamental basado en derechos.

Destinatarios y alcance

19) Estas directrices van dirigidas a todas las partes interesadas implicadas en el diseño, desarrollo, despliegue, aplicación o utilización de IA, o que se vean afectadas por esta, incluidas, con carácter no limitativo, las empresas, organizaciones, investigadores, servicios públicos, agencias gubernamentales, instituciones, organizaciones de la sociedad civil, particulares, trabajadores y consumidores. Las partes interesadas comprometidas con una IA fiable pueden optar voluntariamente por utilizar estas directrices como método para poner en práctica su compromiso, concretamente empleando la lista de evaluación que encontrarán en el capítulo III en los procesos de desarrollo y despliegue de sus sistemas de IA. Esta lista de evaluación también puede complementar —y, por lo tanto, incorporarse a— los procesos de evaluación existentes.

20) Estas directrices pretenden ofrecer orientaciones sobre las aplicaciones de la IA en general y establecer una base horizontal para lograr una IA fiable. A pesar de ello, **las diferentes situaciones plantean desafíos distintos**. Los sistemas de IA de recomendaciones musicales no suscitan las mismas preocupaciones éticas que los que proponen tratamientos médicos críticos. De igual modo, surgen oportunidades y retos diferentes de los sistemas de IA utilizados en el contexto de las relaciones entre empresas y consumidores, entre empresas, entre un empresario y sus trabajadores o entre el sector público y la ciudadanía, o, desde un punto de vista más general, en sectores o casos de uso diversos. Dado que los sistemas de IA son específicos del contexto, los autores de estas directrices reconocen que su uso deberá adaptarse a la aplicación concreta de la inteligencia artificial. Además, debería explorarse la necesidad de adoptar también un enfoque sectorial que complemente el marco horizontal más general propuesto en este documento.

Para comprender mejor cómo pueden aplicarse estas directrices de forma horizontal y los asuntos que requieren un planteamiento sectorial, invitamos a todas las partes interesadas a utilizar con carácter experimental la lista de evaluación de la fiabilidad de la IA (capítulo III) que pone en práctica dicho marco y a hacernos llegar sus comentarios al respecto. A partir dichas observaciones, que se recopilarán a través de esta fase piloto, procederemos a revisar la lista de evaluación de estas directrices a principios de 2020. La fase piloto comenzará en el verano de 2019 y durará hasta finales de año. Todas las partes interesadas podrán participar en ella indicando su interés a través de la Alianza europea de la IA.

B. UN MARCO PARA UNA IA FIABLE

21) Las presentes directrices articulan un marco para lograr una IA fiable basada en los derechos humanos consagrados en la Carta de los Derechos Fundamentales de la Unión Europea (la «Carta de la UE»), así como en la pertinente legislación internacional de derechos humanos. A continuación se describen brevemente los tres componentes de una IA fiable.

IA lícita

22) Los sistemas de IA no operan en un mundo sin leyes. Existen varias normas jurídicamente vinculantes a escala europea, nacional e internacional que ya son aplicables o pertinentes al desarrollo, despliegue y utilización de sistemas de IA. Entre las fuentes jurídicas de importancia en la materia figuran, con carácter no limitativo, el Derecho primario de la UE (los Tratados de la Unión Europea y su Carta de Derechos Fundamentales) y el

Derecho secundario de la Unión (como el Reglamento General de Protección de Datos, las Directivas contra la discriminación, la Directiva sobre máquinas, la Directiva sobre responsabilidad por los daños causados por productos defectuosos, el Reglamento sobre la libre circulación de datos no personales, las leyes de protección de los consumidores y las Directivas relativas a la seguridad y la salud en el trabajo), pero también los tratados de derechos humanos de las Naciones Unidas y los convenios del Consejo de Europa (como el Convenio Europeo de Derechos Humanos), además de numerosas leyes de los Estados miembros de la UE. Además de las normas de aplicación horizontal, existen diversas normas de carácter sectorial aplicables a determinados usos de la IA (como, por ejemplo, el Reglamento sobre productos sanitarios en el sector de la asistencia sanitaria).

23) Las leyes establecen obligaciones tanto positivas como negativas, lo que significa que no deben interpretarse únicamente en referencia a lo que *no* se puede hacer, sino también en referencia a lo que *debe* hacerse. Las leyes no solo prohíben ciertas acciones; también posibilitan otras. En este sentido es preciso señalar que la Carta de la UE contiene artículos sobre la «libertad de empresa» y la «libertad de las artes y de las ciencias», además de otros que abordan ámbitos con los que estamos más familiarizados cuando tratamos de garantizar la fiabilidad de la IA, como, por ejemplo, la protección de datos y la no discriminación.

24) En estas directrices no se trata explícitamente el primer componente de una IA fiable (la IA lícita); en su lugar, se ofrecen orientaciones sobre cómo fomentar y garantizar los otros dos componentes (IA ética y robusta). Si bien estos dos últimos componentes ya están recogidos en cierta medida en las leyes existentes, su plena realización puede trascender las obligaciones legales actualmente existentes.

25) En ningún caso se interpretará que el contenido de este documento pretende ofrecer asesoramiento jurídico u orientaciones acerca de cómo cumplir cualquiera de las normas o requisitos legales existentes. De igual modo, ninguno de los aspectos recogidos en este documento creará derechos ni obligaciones legales para terceros. No obstante, recordamos que cualquier persona física o jurídica tiene el deber de cumplir las leyes actualmente aplicables y aquellas que se adopten en el futuro de acuerdo con la evolución de la inteligencia artificial. Estas directrices parten de la hipótesis de que **todos los derechos y obligaciones legales que se aplican a los procesos y actividades implicados en el desarrollo, despliegue y utilización de la IA conservan su carácter obligatorio y han de ser debidamente observados.**

IA ética

26) Para hacer realidad una IA fiable no solo es necesario cumplir la ley, aunque este es uno de los tres componentes necesarios para ello. Las leyes no siempre avanzan al mismo ritmo que la evolución tecnológica; en ocasiones pueden además ser incoherentes con las normas éticas o, simplemente, poco adecuadas para abordar determinadas cuestiones. En consecuencia, para que los sistemas de AI sean fiables, deben ser también éticos. Esto implica que es preciso garantizar que cumplan las normas éticas.

IA robusta

27) Incluso si se garantiza un fin ético, las personas y la sociedad también deben poder confiar en que los sistemas de IA no provocarán un daño involuntario. Dichos sistemas deben funcionar de manera segura y fiable; se deberían prever medidas de protección para evitar cualquier efecto adverso imprevisto. Así pues, es importante asegurar que los sistemas de IA sean robustos. Esto es necesario tanto desde un punto de vista técnico (garantizando una adecuada solidez técnica del sistema en un contexto determinado, como el ámbito de aplicación o la fase del ciclo de vida en que se utilice), pero también social (teniendo debidamente en cuenta el contexto y el entorno en el que opera el sistema). Una IA ética y robusta, por tanto, están estrechamente interrelacionadas y se complementan entre sí. Los principios que se exponen en el capítulo I y los requisitos que se derivan de ellos y se describen en el capítulo II abordan ambos componentes.

El marco

28) Las directrices recogidas en este documento se ofrecen en tres niveles de abstracción, desde el capítulo I (el

más abstracto) al III (el más concreto):

I) Fundamentos de una IA fiable. En el capítulo I se sientan las bases de una IA fiable, definiendo su enfoque basado en los derechos fundamentales[12]. En dicho capítulo se identifican y describen los principios éticos que deben cumplirse para garantizar una IA ética y robusta.

II) Realización de la IA fiable. El capítulo II traduce los principios éticos anteriores en siete requisitos que los sistemas de IA deben aplicar y cumplir a lo largo de todo su ciclo de vida. Además, ofrece tanto métodos técnicos como de otro tipo que se pueden utilizar para su aplicación.

III) Evaluación de la IA fiable. Los profesionales de la IA esperan recibir orientaciones concretas. Por consiguiente, el capítulo III propone una lista de evaluación preliminar y no exhaustiva de lA fiable con el fin de poner en práctica los requisitos descritos en el capítulo II. Esta evaluación deberá adaptarse a la aplicación específica del sistema de IA.

29) La sección final del documento proporciona ejemplos de oportunidades beneficiosas que se deberían perseguir y de preocupaciones cruciales que plantean los sistemas de IA y que se deberían tener en cuenta, y sobre las que nos gustaría estimular un debate más profundo.

30) La *ilustración 1* muestra la estructura de esta guía.

[12] Los derechos fundamentales se encuentran en la base de la legislación de derechos humanos, tanto a escala internacional como de la UE, y sustentan los derechos legalmente exigibles garantizados por los Tratados de la UE y la Carta de los Derechos Fundamentales de la UE. Dado que los derechos fundamentales son jurídicamente vinculantes, su cumplimiento entra dentro del primero de los componentes de una IA fiable, que hemos denominado la «IA lícita». No obstante, los derechos fundamentales también pueden entenderse como derechos morales especiales de todas las personas por el hecho de serlo, con independencia de su carácter jurídicamente vinculante. En ese sentido, también forman parte del segundo componente de la IA fiable, la «IA ética».

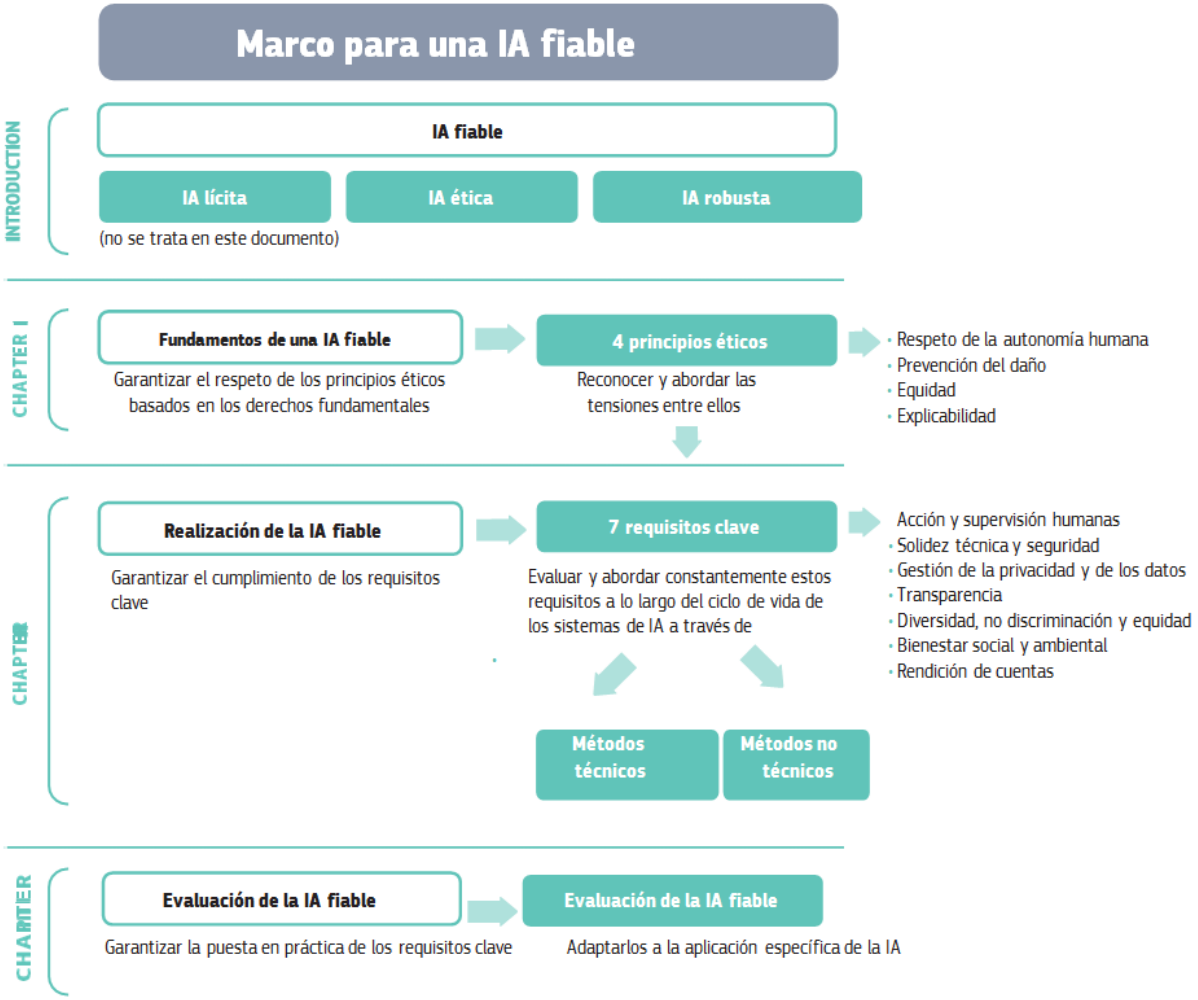

Ilustración 1: Las directrices como marco para una IA fiable

I. Capítulo I: Fundamentos de una IA fiable

31) En este capítulo se sientan las bases de una IA fiable, basada en los derechos fundamentales y respaldada por cuatro principios éticos que deben cumplirse para garantizar una IA ética y robusta. Este capítulo se centra en el campo de la ética.

32) La ética de la inteligencia artificial es un subcampo de la ética aplicada que estudia los problemas éticos que plantea el desarrollo, despliegue y utilización de la IA. Su preocupación fundamental es identificar de qué modo puede la IA mejorar o despertar inquietudes para la vida de las personas, ya sea en términos de calidad de vida o de la autonomía y la libertad humanas necesarias para una sociedad democrática.

33) La reflexión ética sobre la tecnología de la IA puede servir para múltiples fines. En primer lugar, puede estimular la reflexión sobre la necesidad de proteger a las personas y los grupos en el nivel más básico. En segundo lugar, puede estimular nuevos tipos de innovaciones que busquen fomentar valores éticos, como aquellas que contribuyen a lograr los Objetivos de Desarrollo Sostenible de las Naciones Unidas[13], que se encuentran firmemente integradas en la próxima Agenda 2030 de la UE[14]. Pese a que este documento se centra en el primero de estos fines, no se debe subestimar la importancia que puede tener la ética en el segundo de ellos. Una IA fiable puede mejorar el bienestar individual y el bienestar colectivo mediante la generación de prosperidad, la creación de valor y la maximización de la riqueza. Puede contribuir a construir una sociedad justa, ayudando a mejorar la salud y el bienestar de los ciudadanos de maneras que promuevan la igualdad en la distribución de las oportunidades económicas, sociales y políticas.

34) Por lo tanto, es imprescindible que comprendamos cuál es la mejor forma de apoyar el desarrollo, despliegue y utilización de la IA, con objeto de garantizar que cualquier persona pueda prosperar en un mundo basado en la inteligencia artificial, y crear un futuro mejor, manteniendo al mismo tiempo nuestra competitividad a escala mundial. Como sucede con cualquier tecnología potente, el uso de sistemas de IA en nuestra sociedad plantea diversos desafíos éticos, relacionados, por ejemplo, con los efectos de esta tecnología sobre las personas y la sociedad, las capacidades de adopción de decisiones y la seguridad. Si se va a realizar un uso creciente de los sistemas de AI o a delegar decisiones en ellos, es necesario que nos aseguremos de que dichos sistemas afectan de un modo equitativo a la vida de las personas, que están en consonancia con los valores que consideramos fundamentales y que son capaces de actuar en consecuencia. Para ello, necesitamos contar con procesos de rendición de cuentas adecuados.

35) Europa debe definir la visión normativa que desea perseguir sobre un futuro en el que la IA tenga un papel preponderante y, por tanto, entender qué concepto de IA debe estudiarse, desarrollarse, desplegarse y utilizarse en Europa para hacer realidad esa visión. Con este documento, nuestra intención es contribuir a este esfuerzo introduciendo el concepto de IA fiable, que creemos que es la forma adecuada de construir un futuro con IA. Un futuro en el que la democracia, el estado de Derecho y los derechos fundamentales sustenten los sistemas de IA, y en el que dichos sistemas mejoren constantemente y defiendan la cultura democrática, también hará posible un entorno en el que puedan prosperar la innovación y la competitividad responsable.

36) Un código ético específico de un ámbito determinado —por muy coherentes y sofisticadas que puedan ser sus futuras versiones— nunca puede sustituir al razonamiento ético, que debe ser sensible en todo momento a detalles contextuales que no es posible capturar en unas directrices generales. Para garantizar la fiabilidad de la IA es necesario, más allá de desarrollar un conjunto de normas, crear y mantener una cultura y una mentalidad éticas a través del debate público, la educación y el aprendizaje práctico.

1. Los derechos fundamentales como derechos morales y legales

13 https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_es.

14 https://sustainabledevelopment.un.org/?menu=1300.

37) Creemos en un enfoque de la ética en la IA basado en los derechos fundamentales consagrados en los Tratados de la UE[15], la Carta de los Derechos Fundamentales de la Unión Europea (la «Carta de la UE») y la legislación internacional de derechos humanos[16]. El respeto de los derechos fundamentales, dentro de un marco de democracia y estado de Derecho, proporciona la base más prometedora para identificar los principios y valores éticos abstractos que se pueden poner en práctica en el contexto de la IA.

38) Los Tratados de la Unión y la Carta de la UE prescriben una serie de derechos fundamentales que los Estados miembros y las instituciones de la Unión Europea tienen la obligación legal de respetar al aplicar la legislación de la UE. Dichos derechos se describen en la Carta de la UE mediante referencias a la dignidad, las libertades, la igualdad y la solidaridad, los derechos de los ciudadanos y la justicia. La base común a todos estos derechos puede considerarse arraigada en el respeto de la dignidad humana, reflejando así lo que describimos como un «enfoque centrado en la persona» en el que el ser humano disfruta de una condición moral única e inalienable de primacía en las esferas civil, política, económica y social[17].

39) Pese a que los derechos recogidos en la Carta de la UE son jurídicamente vinculantes[18], es importante reconocer que los derechos fundamentales no siempre ofrecen una protección jurídica integral. En lo que respecta a la Carta de la UE, por ejemplo, es importante subrayar que su ámbito de aplicación se limita a áreas del Derecho de la UE. La legislación internacional de derechos humanos y, en particular, el Convenio Europeo de Derechos Humanos son de obligado cumplimiento para los Estados miembros de la UE, incluso en campos que quedan fuera del ámbito de aplicación del Derecho de la UE. Al mismo tiempo, es preciso destacar que los derechos fundamentales también se atribuyen a personas físicas y (en cierta medida) grupos en virtud de su condición moral como seres humanos, con independencia de la fuerza legal de dichos derechos. Entendidos como derechos legalmente exigibles, los derechos fundamentales forman parte, por tanto, del primer componente de la IA fiable (la IA lícita), que garantiza el cumplimiento de la legislación. Entendidos como derechos de cualquier persona, arraigados en la condición moral inherente a los seres humanos, también sustentan el segundo componente de la IA fiable (la IA ética), que se ocupa de las normas éticas que, pese a no ser necesariamente jurídicamente vinculantes, son cruciales para garantizar la fiabilidad. Dado que este documento no tiene el objetivo de ofrecer orientación sobre el primer componente, a efectos de estas directrices no vinculantes, las referencias a los derechos fundamentales se entenderán hechas al segundo.

2. De los derechos fundamentales a los principios éticos

2.1 Los derechos fundamentales como base para una IA fiable

40) Entre el amplio conjunto de derechos indivisibles recogido en la legislación internacional de derechos humanos, los Tratados de la UE y la Carta de la UE, las familias siguientes de derechos fundamentales resultan particularmente aptas para cubrir los sistemas de IA. En determinadas circunstancias, muchos de esos derechos son legalmente exigibles en la UE, por lo que son de obligado cumplimiento desde el punto de vista legal. Pero, incluso una vez lograda la exigibilidad legal de los derechos fundamentales, la reflexión ética puede ayudarnos a comprender el modo en que el desarrollo, despliegue y utilización de la IA pueden afectar a los derechos fundamentales y sus valores subyacentes, y de qué manera pueden contribuir a ofrecer orientaciones más detalladas a la hora de tratar de identificar aquello que *debemos* hacer en lugar de lo que *podemos* hacer (actualmente) con la tecnología.

---

15 La UE se basa en un compromiso constitucional de proteger los derechos fundamentales e indivisibles de los seres humanos, garantizar el respeto del estado de Derecho, fomentar la libertad democrática y promover el bien común. Tales derechos están reflejados en los artículos 2 y 3 del Tratado de la Unión Europea, así como en la Carta de los Derechos Fundamentales de la Unión Europea.

16 Otros instrumentos jurídicos reflejan estos mismos compromisos y profundizan en ellos, como, por ejemplo, la Carta Social Europea del Consejo de Europa o determinadas leyes como el Reglamento General de Protección de Datos de la UE.

17 Es preciso señalar que un compromiso con una IA centrada en la persona y con su anclaje en los derechos humanos requiere unos fundamentos sociales y constitucionales colectivos en los que la libertad individual y el respeto de la dignidad humana sean tanto posibles en la práctica como significativos, en lugar de implicar una concepción indebidamente individualista del ser humano.

18 En virtud del artículo 51 de la Carta, se aplica a las instituciones y Estados miembros de la UE cuando aplican el Derecho de la Unión.

41) ***Respeto de la dignidad humana***. La dignidad humana contiene en sí la idea de que todo ser humano posee un «valor intrínseco» que jamás se debe menoscabar, poner en peligro ni ser objeto de represión por parte de otros (ni de las nuevas tecnologías, como los sistemas de IA).[19] En el contexto de la inteligencia artificial, el respeto de la dignidad humana implica que todas las personas han de ser tratadas con el debido respeto que merecen como *sujetos* morales, y no como simples *objetos* que se pueden filtrar, ordenar, puntuar, dirigir, condicionar o manipular. En consecuencia, los sistemas de IA deben desarrollarse de un modo que respete, proteja y esté al servicio de la integridad física y mental de los seres humanos, el sentimiento de identidad personal y cultural y la satisfacción de sus necesidades esenciales[20].

42) ***Libertad individual***. Los seres humanos deben ser libres para tomar decisiones vitales por sí mismos. Esto implica libertad frente a intromisiones soberanas, pero también requiere la intervención de organizaciones gubernamentales y no gubernamentales para garantizar que los individuos o las personas en riesgo de exclusión disfruten de igualdad de acceso a los beneficios y las oportunidades que ofrece la IA. En el contexto de la inteligencia artificial, la libertad individual exige mitigar la coerción ilegítima (in)directa, las amenazas a la autonomía mental y la salud mental, la vigilancia injustificada, el engaño y la manipulación injusta. De hecho, la libertad individual entraña un compromiso de permitir que los individuos ejerzan un control aún mayor sobre su vida, incluidos (entre otros derechos) la protección de la libertad de empresa, la libertad de las artes y de las ciencias, la libertad de expresión, el derecho a la privacidad y la vida privada y la libertad de reunión y asociación.

43) ***Respeto de la democracia, la justicia y el estado de Derecho***. En las democracias constitucionales, todo poder gubernamental debe estar autorizado legalmente y limitado por la legislación. Los sistemas de IA deberían servir para mantener e impulsar procesos democráticos, así como para respetar la pluralidad de valores y elecciones vitales de las personas. Los sistemas de IA no deben socavar los procesos democráticos, las deliberaciones humanas ni los sistemas democráticos de votación. Asimismo, los sistemas de IA deben incluir un compromiso de garantizar que su funcionamiento no menoscabe los compromisos esenciales en los que se fundamenta el estado de Derecho —así como las leyes y reglamentos de obligado cumplimiento— y de asegurar el respeto de las garantías procesales y la igualdad ante la ley.

44) ***Igualdad, no discriminación y solidaridad****, incluidos los derechos de las personas en riesgo de exclusión.* Es preciso garantizar por igual el respeto del valor moral y la dignidad de todos los seres humanos. Este requisito va más allá de la no discriminación, que tolera el establecimiento de distinciones entre situaciones diferentes sobre la base de justificaciones objetivas. En el contexto de la IA, la igualdad implica que el funcionamiento de este tipo de sistemas no debe generar resultados injustamente sesgados (por ejemplo, los datos utilizados para la formación de los sistemas de IA deben ser lo más inclusivos posibles, de forma que estén representados los diferentes grupos de población). Esto también requiere un adecuado respeto de las personas y grupos potencialmente vulnerables[21], como los trabajadores, las mujeres, las personas con discapacidad, las minorías étnicas, los niños, los consumidores u otras personas en riesgo de exclusión.

45) ***Derechos de los ciudadanos***. Los ciudadanos disfrutan de una amplia variedad de derechos, como el derecho de voto, el derecho a una buena administración, el derecho de acceso a documentos públicos o el derecho de petición a la administración. Los sistemas de IA ofrecen un potencial muy importante para mejorar el alcance y la eficiencia del gobierno en la prestación de bienes y servicios públicos a la sociedad. Al mismo tiempo, determinadas aplicaciones de la IA también pueden afectar negativamente a los derechos de los ciudadanos, que deben protegerse. La utilización del término «derechos de los ciudadanos» en el presente documento no significa que se nieguen o ignoren los derechos de los nacionales de terceros países o de las personas que se

---

19 C. McCrudden, *Human Dignity and Judicial Interpretation of Human Rights*, *EJIL*, 19(4), 2008.

20 Para comprender el concepto de «dignidad humana» utilizado en este documento, véase E. Hilgendorf, «Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity», en: D. Grimm, A. Kemmerer, C. Möllers (eds.), *Human Dignity in Context. Explorations of a Contested Concept*, 2018, pp. 325 y ss.

21 En el glosario puede consultarse una descripción del término tal como se utiliza a lo largo de este documento.

encuentren en situación irregular (o ilegal) en la UE, que también tienen derechos al amparo de la legislación internacional, incluso —por tanto— en el campo de la IA.

2.2 Principios éticos en el contexto de los sistemas de IA[22]

46) Muchas organizaciones públicas, privadas y civiles se han inspirado en los derechos fundamentales para elaborar marcos éticos para la IA[23]. En la UE, el Grupo europeo de ética de la ciencia y de las nuevas tecnologías propuso un conjunto de nueve principios básicos, apoyados en los valores fundamentales recogidos en los Tratados de la UE y en la Carta de los Derechos Fundamentales de la Unión Europea[24]. Este documento también se basa en ese trabajo, reconociendo la mayoría de los principios que propugnan los diferentes grupos y, al mismo tiempo, aclarando los fines que todos esos principios tratan de alimentar y respaldar. Estos principios éticos pueden inspirar instrumentos reglamentarios nuevos y específicos, contribuir a interpretar los derechos fundamentales a medida que vaya evolucionando nuestro entorno sociotécnico y guiar la lógica del desarrollo, utilización y aplicación de los sistemas de IA, de forma que se adapten dinámicamente conforme evolucione la propia sociedad.

47) Los sistemas de IA deberían mejorar el bienestar individual y colectivo. En esta sección se enumeran **cuatro principios éticos**, arraigados en los derechos fundamentales, que deben cumplirse para garantizar que los sistemas de IA se desarrollen, despliegen y utilicen de manera fiable. Se establece específicamente que se trata de **imperativos éticos** que los profesionales de la IA deben esforzarse en todo momento por observar. Sin imponer una jerarquía entre ellos, los principios se enumeran a continuación siguiendo el orden de aparición de los derechos fundamentales en los que se basan en la Carta de la UE[25].

48) Se trata de los principios de:

I) respeto de la autonomía humana;
II) prevención del daño;
III) equidad;
IV) explicabilidad.

49) Muchos de ellos aparecen recogidos ya en gran medida en los requisitos legales existentes de obligado cumplimiento, por lo que entran dentro del ámbito de la «IA lícita», el primer componente de la IA fiable[26]. Sin embargo, como se ha señalado anteriormente, pese a que numerosas obligaciones legales reflejan principios éticos, el cumplimiento de estos últimos trasciende el mero cumplimiento de las leyes existentes[27].

- <u>El principio del respeto de la autonomía humana</u>

---

22 Estos principios se aplican también al desarrollo, despliegue y utilización de otras tecnologías, por lo que no son específicos de los sistemas de IA. En las páginas que siguen hemos pretendido establecer su pertinencia en un contexto específicamente relacionado con la IA.

23 El hecho de basarnos en los derechos fundamentales ayuda también a limitar la incertidumbre reglamentaria, pues contamos con décadas de práctica en la protección de tales derechos en la UE que nos ofrecen claridad, legibilidad y previsibilidad.

24 Más recientemente, el grupo de trabajo AI4People ha estudiado los mencionados principios del Grupo europeo de ética de la ciencia y de las nuevas tecnologías y otros treinta y seis principios éticos propuestos hasta la fecha, y los ha aglutinado en cuatro principios generales: L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), "AI4People —An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", Minds and Machines 28(4): 689-707.

25 El respeto de la autonomía humana está estrechamente relacionado con el derecho a la dignidad y la libertad humanas (recogido en los artículos 1 y 6 de la Carta). La prevención del daño está fuertemente vinculada a la protección de la integridad física o mental (reflejada en el artículo 3). La equidad está estrechamente asociada a los derechos a la no discriminación, la solidaridad y la justicia (recogidos en el artículo 21 y siguientes). La explicabilidad y la responsabilidad están relacionadas, a su vez, con los derechos referentes a la justicia (reflejados en el artículo 47).

26 Piénsese, por ejemplo, en el Reglamento General de Protección de Datos o en los reglamentos de la UE en materia de protección de los consumidores.

27 Para obtener información adicional sobre este tema, consúltese, por ejemplo, L. Floridi, «Soft Ethics and the Governance of the Digital», *Philosophy & Technology*, marzo de 2018, volumen 31, n.º 1, pp. 1–8.

50) Los derechos fundamentales en los que se apoya la UE van dirigidos a garantizar el respeto de la libertad y la autonomía de los seres humanos. Las personas que interactúen con sistemas de IA deben poder mantener una autonomía plena y efectiva sobre sí mismas y ser capaces de participar en el proceso democrático. Los sistemas de IA no deberían subordinar, coaccionar, engañar, manipular, condicionar o dirigir a los seres humanos de manera injustificada. En lugar de ello, los sistemas de IA deberían diseñarse de forma que aumenten, complementen y potencien las aptitudes cognitivas, sociales y culturales de las personas. La distribución de funciones entre los seres humanos y los sistemas de IA debería seguir principios de diseño centrados en las personas, y dejar amplias oportunidades para la elección humana. Esto implica garantizar la supervisión[28] y el control humanos sobre los procesos de trabajo de los sistemas de IA. Los sistemas de IA también pueden transformar de un modo fundamental el mundo del trabajo. Deberían ayudar a las personas en el entorno laboral y aspirar a crear empleos útiles.

- El principio de prevención del daño

51) Los sistemas de IA no deberían provocar daños (o agravar los existentes)[29] ni perjudicar de cualquier otro modo a los seres humanos[30]. Esto conlleva la protección de la dignidad humana, así como de la integridad física y mental. Todos los sistemas y entornos de IA en los que operan estos deben ser seguros. También deberán ser robustos desde el punto de vista técnico, y debería garantizarse que no puedan destinarse a usos malintencionados. Las personas vulnerables deberían recibir mayor atención y participar en el desarrollo y despliegue de los sistemas de IA. Se deberá prestar también una atención particular a las situaciones en las que los sistemas de IA puedan provocar efectos adversos (o agravar los existentes) debido a asimetrías de poder o de información, por ejemplo entre empresarios y trabajadores, entre empresas y consumidores o entre gobiernos y ciudadanos. La prevención del daño implica asimismo tener en cuenta el entorno natural y a todos los seres vivos.

- El principio de equidad

52) El desarrollo, despliegue y utilización de sistemas de IA debe ser equitativo. Pese a que reconocemos que existen muchas interpretaciones diferentes de la equidad, creemos que esta tiene tanto una dimensión sustantiva como procedimental. La dimensión sustantiva implica un compromiso de: garantizar una distribución justa e igualitaria de los beneficios y costes, y asegurar que las personas y grupos no sufran sesgos injustos, discriminación ni estigmatización. Si se pueden evitar los sesgos injustos, los sistemas de IA podrían incluso aumentar la equidad social. También se debería fomentar la igualdad de oportunidades en términos de acceso a la educación, los bienes los servicios y la tecnología. Además, el uso de sistemas de IA no debería conducir jamás a que se engañe a los usuarios (finales) ni se limite su libertad de elección. Asimismo, la equidad implica que los profesionales de la IA deberían respetar el principio de proporcionalidad entre medios y fines, y estudiar cuidadosamente cómo alcanzar un equilibrio entre los diferentes intereses y objetivos contrapuestos[31]. La dimensión procedimental de la equidad conlleva la capacidad de oponerse a las decisiones adoptadas por los sistemas de IA y por las personas que los manejan, así como de tratar de obtener

28 El concepto de supervisión humana se desarrolla en el punto 65.

29 Los daños pueden ser individuales o colectivos, e incluir daños intangibles al entorno social, cultural y político.

30 Esto también abarca el modo de vida de los individuos y grupos sociales, evitando causar, por ejemplo, daños culturales.

31 Esto guarda relación con el principio de proporcionalidad (reflejado en la máxima de «no matar moscas a cañonazos»). Las medidas adoptadas para lograr un fin (por ejemplo, las medidas de extracción de datos que se introduzcan para que la IA cumpla su función de optimización) deberían limitarse a las estrictamente necesarias. Esto implica además que, cuando varias medidas compitan por la consecución de un objetivo, debería darse prioridad a aquella que menos perjudique los derechos fundamentales y las normas éticas (por ejemplo, los desarrolladores de IA deberían preferir en todo momento datos proporcionados por el sector público frente a los datos personales). También cabe hacer referencia a la proporcionalidad entre el usuario y el responsable del despliegue, teniendo en cuenta, por una parte, los derechos de las empresas (incluida la propiedad intelectual y la confidencialidad) y, por otra, los derechos de los usuarios.

compensaciones adecuadas frente a ellas[32]. Con este fin, se debe poder identificar a la entidad responsable de la decisión y explicar los procesos de adopción de decisiones.

- El principio de explicabilidad

53) La explicabilidad es crucial para conseguir que los usuarios confíen en los sistemas de IA y para mantener dicha confianza. Esto significa que los procesos han de ser transparentes, que es preciso comunicar abiertamente las capacidades y la finalidad de los sistemas de IA y que las decisiones deben poder explicarse —en la medida de lo posible— a las partes que se vean afectadas por ellas de manera directa o indirecta. Sin esta información, no es posible impugnar adecuadamente una decisión. No siempre resulta posible explicar por qué un modelo ha generado un resultado o una decisión particular (ni qué combinación de factores contribuyeron a ello). Esos casos, que se denominan algoritmos de «caja negra», requieren especial atención. En tales circunstancias, puede ser necesario adoptar otras medidas relacionadas con la explicabilidad (por ejemplo, la trazabilidad, la auditabilidad y la comunicación transparente sobre las prestaciones del sistema), siempre y cuando el sistema en su conjunto respete los derechos fundamentales. El grado de necesidad de explicabilidad depende en gran medida del contexto y la gravedad de las consecuencias derivadas de un resultado erróneo o inadecuado[33].

2.3 Tensiones entre los diferentes principios

54) Cabe la posibilidad de que surjan tensiones entre los principios anteriores, y no existe una solución establecida para resolverlas. En consonancia con el compromiso fundamental de la UE con la participación democrática, el respeto de las garantías procesales y la participación abierta en la esfera política, deberían establecerse métodos que posibiliten un debate responsable sobre dichas tensiones. A modo de ejemplo, los principios de *prevención del daño* y de *autonomía humana* pueden entrar en conflicto en diversos ámbitos. Considérese el ejemplo de la utilización de sistemas de IA para la «actuación policial predictiva», que puede ayudar a reducir la delincuencia, pero de formas que incluyan actividades de vigilancia que vulneren la libertad y la privacidad individuales. Además, los beneficios globales de los sistemas de IA deberían ser sustancialmente superiores a los riesgos individuales previsibles. Pese a que estos principios ofrecen ciertamente una orientación para la búsqueda de soluciones, no dejan de ser prescripciones éticas abstractas. Por lo tanto, no se debe esperar que los profesionales de la IA encuentren la solución adecuada basándose en los principios anteriores; no obstante, deberán afrontar los dilemas éticos y analizar las ventajas e inconvenientes a través de un proceso de reflexión razonada con base empírica, en lugar de guiarse por la intuición o por criterios aleatorios. Pese a todo, pueden existir situaciones en las que no sea posible identificar compensaciones aceptables desde el punto de vista ético. Determinados derechos fundamentales y principios correlacionados son de carácter absoluto y no pueden ser objeto de un ejercicio de búsqueda de equilibrio (es el caso, por ejemplo, de la dignidad humana).

[32] Incluso mediante el ejercicio de su derecho de asociación y de afiliación a un sindicato en un entorno laboral, según lo dispuesto en el artículo 12 de la Carta de los Derechos Fundamentales de la Unión Europea.

[33] Por ejemplo, un sistema de IA que genere unas recomendaciones de compra poco acertadas no despertará excesivas preocupaciones desde el punto de vista ético, a diferencia de los sistemas de IA que evalúan si se debería conceder la libertad condicional a una persona condenada por un delito penal.

Orientaciones clave derivadas del capítulo I:

- ✓ Desarrollar, desplegar y utilizar los sistemas de IA respetando los principios éticos de: *respeto de la autonomía humana, prevención del daño, equidad y explicabilidad.* Reconocer y abordar las tensiones que pueden surgir entre estos principios.
- ✓ Prestar una atención especial a las situaciones que afecten a los grupos más vulnerables, como los niños, las personas con discapacidad y otros colectivos que se hayan visto históricamente desfavorecidos o que se encuentren en riesgo de exclusión o en situaciones caracterizadas por asimetrías de poder o de información, como las que pueden producirse entre empresarios y trabajadores o entre empresas y consumidores[34].
- ✓ Reconocer y tener presente que, pese a que pueden aportar numerosos y sustanciales beneficios a las personas y a la sociedad, algunas aplicaciones de IA también pueden tener efectos negativos, algunos de los cuales pueden resultar difíciles de prever, identificar o medir (por ejemplo, sobre la democracia, el estado de Derecho y la justicia distributiva, o sobre la propia mente humana). Adoptar medidas adecuadas para mitigar estos riesgos cuando proceda; dichas medidas deberán ser proporcionales a la magnitud del riesgo.

II. Capítulo II: Realización de la IA fiable

55) Este capítulo ofrece orientaciones de cara a la materialización y el logro de una IA fiable a través de una lista de siete requisitos que se deberían cumplir, basados en los principios establecidos en el capítulo I. Además, se presenta una serie de métodos técnicos y no técnicos actualmente disponibles para garantizar el cumplimiento de dichos requisitos a lo largo de todo el ciclo de vida de los sistemas de IA.

1. Requisitos de una IA fiable

56) Los principios expuestos en el capítulo I deben traducirse en requisitos concretos para hacer realidad una IA fiable. Dichos requisitos son aplicables a las diferentes partes interesadas que participan en algún momento del ciclo de vida de los sistemas de IA: desarrolladores, responsables del despliegue y usuarios finales, así como a la sociedad en su conjunto. Con el término «desarrolladores» nos referimos a las personas dedicadas a la investigación, el diseño o el desarrollo de sistemas de IA. Por «responsables del despliegue» entendemos las organizaciones públicas o privadas que utilizan sistemas de IA en sus procesos internos y para ofrecer productos y servicios a otros agentes. Los «usuarios finales» son aquellos que interactúan con el sistema de IA, ya sea de forma directa o indirecta. Por último, la «sociedad en su conjunto» engloba el resto de agentes, personas y entidades afectados de manera directa o indirecta por los sistemas de IA.

57) Las diferentes clases de partes interesadas tienen diversos papeles que desempeñar para garantizar el cumplimiento de los requisitos:

a. los desarrolladores deben introducir y aplicar los requisitos de los procesos de diseño y desarrollo;
b. los responsables del despliegue deben asegurarse de que los sistemas que utilizan y los productos y servicios que ofrecen cumplen los requisitos establecidos;
c. los usuarios finales y la sociedad en su conjunto deben permanecer informados sobre dichos requisitos y tener la capacidad de pedir que se cumplan.

[34] Véanse los artículos 24 a 27 de la Carta de la UE, que tratan sobre los derechos de los niños y de las personas mayores, la integración de las personas con discapacidad y los derechos de los trabajadores. Véase también el artículo 38 relativo a la protección de los consumidores.

58) A continuación se ofrece una lista no exhaustiva de los requisitos[35]. Incluye aspectos sistémicos, individuales y sociales:

1 **Acción y supervisión humanas**
Incluidos los derechos fundamentales, la acción humana y la supervisión humana.

2 **Solidez técnica y seguridad**
Incluida la capacidad de resistencia a los ataques y la seguridad, un plan de repliegue y la seguridad general, precisión, fiabilidad y reproducibilidad.

3 **Gestión de la privacidad y de los datos**
Incluido el respeto de la privacidad, la calidad y la integridad de los datos, así como el acceso a estos.

4 **Transparencia**
Incluidas la trazabilidad, la explicabilidad y la comunicación.

5 **Diversidad, no discriminación y equidad**
Incluida la ausencia de sesgos injustos, la accesibilidad y el diseño universal, así como la participación de las partes interesadas.

6 **Bienestar social y ambiental**
Incluida la sostenibilidad y el respeto del medio ambiente, el impacto social, la sociedad y la democracia.

7 **Rendición de cuentas**
Incluidas la auditabilidad, la minimización de efectos negativos y la notificación de estos, la búsqueda de equilibrios y las compensaciones.

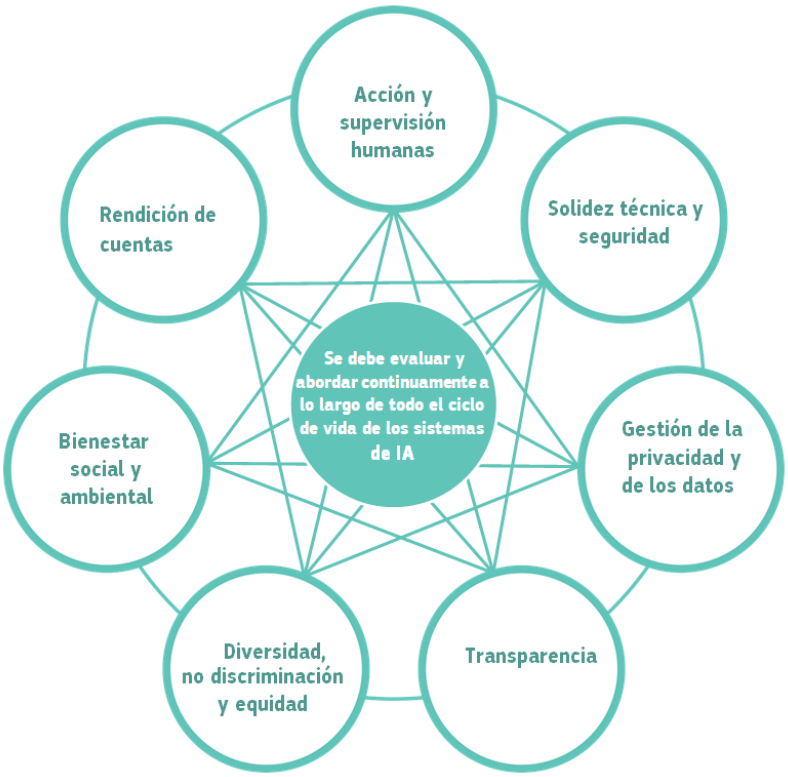

Ilustración 2: Interrelaciones existentes entre los siete requisitos: todos tienen idéntica importancia, se apoyan entre sí

[35] Sin imponer una jerarquía entre ellos, los principios se enumeran a continuación siguiendo el orden de aparición de los principios y derechos con los que están relacionados en la Carta de la UE.

59) Pese a que todos los requisitos tienen la misma importancia, será necesario tener en cuenta el contexto y las tensiones que pueden surgir entre ellos a la hora de aplicarlos en diferentes ámbitos y sectores. Estos requisitos deberían satisfacerse a lo largo de todo el ciclo de vida de un sistema de IA, un cumplimiento que depende de la aplicación específica del sistema. Pese a que la mayoría de ellos son aplicables a todos los sistemas de IA, se presta una atención especial a los que afectan de manera directa o indirecta a las personas. Por lo tanto, su pertinencia puede ser menor en el caso de determinadas aplicaciones (en entornos industriales, por ejemplo).

60) Los requisitos anteriores incluyen elementos que, en algunos casos, ya figuran reflejados en las leyes existentes. Reiteramos que —en consonancia con el primer componente de la IA— los desarrolladores y responsables del despliegue de sistemas de IA tienen el deber de garantizar que dichos sistemas cumplan las obligaciones legales vigentes, tanto en lo que respecta a las normas de aplicación horizontal como a la normativa de carácter sectorial.

61) En los párrafos que siguen se analiza con detalle cada uno de los requisitos.

1. Acción y supervisión humanas

62) Los sistemas de IA deberían respaldar la autonomía y la toma de decisiones de las personas, tal como prescribe el principio del *respeto de la autonomía humana*. Esto requiere que los sistemas de IA actúen tanto como facilitadores de una sociedad democrática, próspera y equitativa, apoyando la acción humana y promoviendo los derechos fundamentales, además de permitir la supervisión humana.

63) ***Derechos fundamentales***. Al igual que muchas otras tecnologías, los sistemas de IA pueden tanto favorecer los derechos fundamentales como obstaculizarlos. Estos sistemas pueden ser beneficiosos para las personas, por ejemplo ayudándolas a llevar a cabo un seguimiento de sus datos personales o mejorando la accesibilidad de la educación, facilitando así el ejercicio del derecho a la educación. Sin embargo, dado el alcance y la capacidad de los sistemas de IA, también pueden afectar negativamente a los derechos fundamentales. En situaciones en las que existan riesgos de este tipo, deberá llevarse a cabo una evaluación del impacto sobre los derechos fundamentales. Esta evaluación debería llevarse a cabo antes del desarrollo de los sistemas de IA en cuestión e incluir una evaluación de las posibilidades de reducir dichos riesgos o de justificar estos como necesarios en una sociedad democrática para respetar los derechos y libertades de otras personas. Además, deberían crearse mecanismos que permitan conocer las opiniones externas sobre los sistemas de IA que pueden vulnerar los derechos fundamentales.

64) ***Acción humana.*** Los usuarios deberían ser capaces de tomar decisiones autónomas con conocimiento de causa en relación con los sistemas de IA. Se les deberían proporcionar los conocimientos y herramientas necesarios para comprender los sistemas de IA e interactuar con ellos de manera satisfactoria y, siempre que resulte posible, permitírseles evaluar por sí mismos o cuestionar el sistema. Los sistemas de IA deberían ayudar a las personas a tomar mejores decisiones y con mayor conocimiento de causa de conformidad con sus objetivos. En ocasiones se pueden desplegar sistemas de IA con el objetivo de condicionar e influir en el comportamiento humano a través de mecanismos que pueden ser difíciles de detectar, dado que pueden explotar procesos del subconsciente mediante diversas formas de manipulación injusta, engaño, dirección y condicionamiento, todas las cuales pueden suponer una amenaza para la autonomía individual. El principio general de autonomía del usuario debe ocupar un lugar central en la funcionalidad del sistema. La clave para ello es el derecho a no ser sometido a una decisión basada exclusivamente en procesos automatizados cuando tal decisión produzca efectos jurídicos sobre los usuarios o les afecte de forma significativa por motivos

similares[36].

65) ***Supervisión humana***. La supervisión humana ayuda a garantizar que un sistema de IA no socave la autonomía humana o provoque otros efectos adversos. La supervisión se puede llevar a cabo a través de mecanismos de gobernanza, tales como los enfoques de participación humana, control humano o mando humano. La participación humana hace referencia a la capacidad de que intervengan seres humanos en todos los ciclos de decisión del sistema, algo que en muchos casos no es posible ni deseable. El control humano se refiere a la capacidad de que intervengan seres humanos durante el ciclo de diseño del sistema y en el seguimiento de su funcionamiento. Por último, el mando humano es la capacidad de supervisar la actividad global del sistema de IA (incluidos, desde un punto de vista más amplio, sus efectos económicos, sociales, jurídicos y éticos), así como la capacidad de decidir cómo y cuándo utilizar el sistema en una situación determinada. Esto puede incluir la decisión de no utilizar un sistema de IA en una situación particular, establecer niveles de discrecionalidad humana durante el uso del sistema o garantizar la posibilidad de ignorar una decisión adoptada por un sistema. Además, se debe garantizar que los responsables públicos puedan ejercer la supervisión en consonancia con sus respectivos mandatos. Puede ser necesario introducir mecanismos de supervisión en diferentes grados para respaldar otras medidas de seguridad y control, dependiendo del ámbito de aplicación y el riesgo potencial del sistema de IA. Si el resto de las circunstancias no cambian, cuanto menor sea el nivel de supervisión que pueda ejercer una persona sobre un sistema de IA, mayores y más exigentes serán las verificaciones y la gobernanza necesarias.

2. Solidez técnica y seguridad

66) Un componente crucial de la IA fiable es la solidez técnica, que está estrechamente vinculada al *principio de prevención del daño*. La solidez técnica requiere que los sistemas de IA se desarrollen con un enfoque preventivo en relación con los riesgos, de modo que se comporten siempre según lo esperado y minimicen los daños involuntarios e imprevistos, evitando asimismo causar daños inaceptables. Lo anterior debería aplicarse también a los cambios potenciales en su entorno operativo o a la presencia de otros agentes (humanos y artificiales) que puedan interactuar con el sistema de manera contenciosa. Además, debería garantizarse la integridad física y mental de los seres humanos.

67) ***Resistencia a los ataques y seguridad***. Los sistemas de IA, como todos los sistemas de software, deben protegerse frente a las vulnerabilidades que puedan permitir su explotación por parte de agentes malintencionados, como, por ejemplo, los piratas informáticos. Los ataques pueden ir dirigidos contra los datos (envenenamiento de los datos), el modelo (fallo del modelo) o la infraestructura informática subyacente, tanto el software como el hardware. En el caso de que un sistema de IA sea objeto de un ataque, por ejemplo por parte de agentes malintencionados, se podrían alterar los datos y el comportamiento del sistema, de modo que este adopte decisiones diferentes o, sencillamente, se desconecte. Los sistemas y los datos también pueden corromperse debido a intenciones maliciosas o por verse expuestos a situaciones inesperadas. Unos procesos de seguridad insuficientes también pueden dar lugar a decisiones erróneas o incluso a daños físicos. Para que los sistemas de IA se consideren seguros,[37] es preciso tener en cuenta las posibles aplicaciones imprevistas de la IA (aplicaciones que puedan utilizarse para fines diferentes, por ejemplo) así como el abuso potencial de un sistema de IA por parte de agentes malintencionados; también se deberán adoptar medidas para prevenir y mitigar esos riesgos.[38]

---

36 Cabe hacer referencia al artículo 22 del RGPD, en el que ya está recogido este derecho.

37 Véanse, por ejemplo, las consideraciones recogidas en el punto 2.7 del plan coordinado de la Unión Europea sobre la inteligencia artificial.

38 En lo que respecta a la seguridad de los sistemas de IA, puede resultar imprescindible tener la capacidad de crear un círculo virtuoso en el ámbito de la investigación y el desarrollo entre la comprensión de los ataques, el desarrollo de medidas de protección adecuadas y la mejora de las metodologías de evaluación. Para ello, se debería promover la convergencia entre la comunidad dedicada a la IA y la comunidad especializada en seguridad. Además, todos los agentes involucrados tienen la responsabilidad de crear normas comunes

68) ***Plan de repliegue y seguridad general***. Los sistemas de IA deberían contar con salvaguardias que posibiliten un plan de repliegue en el caso de que surjan problemas. Esto puede significar que los sistemas de IA pasen de un procedimiento basado en estadísticas a otro basado en normas, o que soliciten la intervención de un operador humano antes de proseguir con sus actuaciones.[39] Es preciso garantizar que el sistema se comportará de acuerdo con lo que se espera de él sin causar daños a los seres vivos ni al medio ambiente. Esto incluye la minimización de las consecuencias y errores imprevistos. Además, se deberían establecer procesos dirigidos a aclarar y evaluar los posibles riesgos asociados con el uso de sistemas de IA en los diversos ámbitos de aplicación. El nivel de las medidas de seguridad requeridas depende de la magnitud del riesgo que plantee un sistema de IA, que a su vez depende de las capacidades del sistema. Cuando se prevea que el proceso de desarrollo o el propio sistema planteará riesgos particularmente altos, es crucial desarrollar y probar medidas de seguridad de forma proactiva.

69) ***Precisión.*** La precisión está relacionada con la capacidad de un sistema de IA para realizar juicios correctos —como, por ejemplo, clasificar correctamente información en las categorías adecuadas—, o con su capacidad para efectuar predicciones, formular recomendaciones o tomar decisiones correctas basándose en datos o modelos. Un proceso de desarrollo y evaluación explícito y correctamente diseñado puede respaldar, mitigar y corregir los riesgos imprevistos asociados a predicciones incorrectas. Cuando no sea posible evitar este tipo de predicciones, es importante que el sistema pueda indicar la probabilidad de que se produzcan esos errores. Un alto nivel de precisión resulta particularmente crucial en situaciones en que un sistema de IA afecte de manera directa a la vida humana.

70) ***Fiabilidad y reproducibilidad.*** Es esencial que los resultados de los sistemas de IA sean reproducibles, además de fiables. Un sistema de IA fiable es aquel que funciona adecuadamente con un conjunto de información y en diversas situaciones. Esto es necesario para evaluar un sistema de IA y evitar que provoque daños involuntarios. La reproducibilidad describe si un experimento con IA muestra el mismo comportamiento cuando se repite varias veces en las mismas condiciones. Esto permite a los científicos y responsables políticos describir con exactitud lo que hacen los sistemas de IA. Los archivos de replicación[40] pueden facilitar el proceso de ensayo y reproducción de comportamientos.

3. Gestión de la privacidad y de los datos

71) La privacidad es un derecho fundamental que se ve especialmente afectado por los sistemas de IA, y que guarda una estrecha relación con el *principio de prevención del daño*. La prevención del daño a la privacidad también requiere una adecuada gestión de los datos, que abarque la calidad y la integridad de los datos utilizados, su pertinencia en contraste con el ámbito en el que se desplegarán los sistemas de IA, sus protocolos de acceso y la capacidad para procesar datos sin vulnerar la privacidad.

72) ***Protección de la intimidad y de los datos***. Los sistemas de IA deben garantizar la protección de la intimidad y de los datos a lo largo de todo el ciclo de vida de un sistema[41]. Esto incluye la información inicialmente facilitada por el usuario, así como la información generada sobre el usuario en el contexto de su interacción con el sistema (por ejemplo, los productos que genere el sistema de AI para determinados usuarios o la respuesta de estos a determinadas recomendaciones). Los registros digitales del comportamiento humano pueden posibilitar que los sistemas de IA no solo infieran las preferencias de las personas, sino también su orientación sexual, edad, género u opiniones políticas y religiosas. Para permitir que los individuos confíen en

---

en materia de seguridad y protección transfronterizas, así como de establecer un entorno de confianza mutua e impulsar la colaboración internacional. Sobre las posibles medidas, véase *Malicious Use of AI* (Avin S., Brundage M. *et al*., 2018).

39 También se deberán estudiar los posibles escenarios en los que no sería posible contar con intervención humana de manera inmediata.

40 Este término se refiere a los archivos que replicarán cada paso del proceso de desarrollo de un sistema de IA, desde la investigación y la recogida inicial de datos hasta los resultados.

41 Cabe hacer referencia a las leyes existentes en materia de protección de la privacidad, como el RGPD o el próximo Reglamento sobre la privacidad y las comunicaciones electrónicas.

el proceso de recopilación de datos, es preciso garantizar que la información recabada sobre ellos no se utilizará para discriminarlos de forma injusta o ilegal.

73) ***Calidad e integridad de los datos***. La calidad de los conjuntos de datos utilizados es primordial para el desempeño de los sistemas de IA. Cuando se recopilan datos, estos pueden contener sesgos sociales, imprecisiones y errores. Este problema debe abordarse antes de llevar a cabo cualquier tipo de formación en la que se utilice cualquier conjunto de datos. Además, es necesario garantizar la integridad de los datos. La introducción de datos malintencionados en un sistema de IA puede alterar su comportamiento, sobre todo si se trata de sistemas con capacidad de autoaprendizaje. Los procesos y conjuntos de datos utilizados deben ponerse a prueba y documentarse en cada paso, por ejemplo, en la planificación, formación, verificación y despliegue. Esto debería aplicarse igualmente a los sistemas de IA que no hayan sido desarrollados internamente, sino adquiridos externamente.

74) ***Acceso a los datos***. En cualquier organización que maneje datos personales (con independencia de si alguien es usuario del sistema o no) deberían establecerse protocolos que rijan el acceso a los datos. En esos protocolos debería describirse quién puede acceder a los datos y en qué circunstancias. Solamente debería permitirse acceder a los datos personales a personal debidamente cualificado, poseedor de las competencias adecuadas y que necesite acceder a la información pertinente.

4. Transparencia

75) Este requisito guarda una relación estrecha con el *principio de explicabilidad* e incluye la transparencia de los elementos pertinentes para un sistema de IA: los datos, el sistema y los modelos de negocio.

76) ***Trazabilidad***. Los conjuntos de datos y los procesos que dan lugar a la decisión del sistema de IA, incluidos los relativos a la recopilación y etiquetado de los datos así como a los algoritmos utilizados, deberían documentarse con arreglo a la norma más rigurosa posible con el fin de posibilitar la trazabilidad y aumentar la transparencia. Esto también es aplicable a las decisiones que adopte el sistema de IA. Esto permitirá identificar los motivos de una decisión errónea por parte del sistema, lo que a su vez podría ayudar a prevenir futuros errores. La trazabilidad, por tanto, facilita la auditabilidad y la explicabilidad.

77) ***Explicabilidad.*** La explicabilidad concierne a la capacidad de explicar tanto los procesos técnicos de un sistema de IA como las decisiones humanas asociadas (por ejemplo, las áreas de aplicación de un sistema de IA). La explicabilidad técnica requiere que las decisiones que adopte un sistema de IA sean comprensibles para los seres humanos y estos tengan la posibilidad de rastrearlas. Además, puede que sea necesario buscar un equilibrio entre la mejora de la explicabilidad de un sistema (que puede reducir su precisión) o una mayor precisión de este (a costa de la explicabilidad). Cuando un sistema de IA tenga un impacto significativo en la vida de las personas, debería ser posible reclamar una explicación adecuada del proceso de toma de decisiones del sistema de IA. Dicha explicación debería ser oportuna y adaptarse al nivel de especialización de la parte interesada (que puede ser una persona no experta en la materia, un regulador o un investigador). Además, debería ser posible disponer de explicaciones sobre la medida en que el sistema de IA condiciona e influye en el proceso de toma de decisiones de la organización, sobre las decisiones de diseño del sistema y sobre la lógica subyacente a su despliegue (garantizando así la transparencia del modelo de negocio).

78) ***Comunicación.*** Los sistemas de IA no deberían presentarse a sí mismos como humanos ante los usuarios; las personas tienen derecho a saber que están interactuando con un sistema de IA. Por lo tanto, los sistemas de IA deben ser identificables como tales. Además, cuando sea necesario, se debería ofrecer al usuario la posibilidad de decidir si prefiere interactuar con un sistema de IA o con otra persona, con el fin de garantizar el cumplimiento de los derechos fundamentales. Más allá de lo expuesto, se debería informar sobre las capacidades y limitaciones del sistema de IA a los profesionales o usuarios finales; dicha información debería proporcionarse de un modo adecuado según el caso de uso de que se trate y debería incluir información acerca del nivel de precisión del sistema de IA, así como de sus limitaciones.

5. Diversidad, no discriminación y equidad

79) Para hacer realidad la IA fiable, es preciso garantizar la inclusión y la diversidad a lo largo de todo el ciclo de vida de los sistemas de inteligencia artificial. Además de tener en cuenta a todos los afectados y garantizar su participación en todo el proceso, también es necesario garantizar la igualdad de acceso mediante procesos de diseño inclusivos, sin olvidar la igualdad de trato. Este requisito está estrechamente relacionado con el *principio de equidad*.

80) ***Necesidad de evitar sesgos injustos***. Los conjuntos de datos que utilizan los sistemas de IA (tanto con fines de formación como para su funcionamiento) pueden presentar sesgos históricos inadvertidos, lagunas o modelos de gestión incorrectos. El mantenimiento de dichos sesgos podría dar lugar a prejuicios y discriminación[42] (in)directos e involuntarios contra determinados grupos o personas, lo que podría agravar los estereotipos y la marginación. La explotación intencionada de los sesgos (de los consumidores) o la competencia desleal también pueden provocar situaciones perjudiciales, como la homogeneización de los precios mediante la colusión o la falta de transparencia del mercado.[43] Siempre que sea posible, los sesgos identificables y discriminatorios deberían eliminarse en la fase de recopilación de la información. Los propios métodos de desarrollo de los sistemas de IA (por ejemplo, la programación de algoritmos) también pueden presentar sesgos injustos. Esto se puede combatir mediante procesos de supervisión que permitan analizar y abordar el propósito, las restricciones, los requisitos y las decisiones del sistema de un modo claro y transparente. Además, la contratación de personas procedentes de diversos contextos, culturas y disciplinas puede garantizar la diversidad de opiniones y debería fomentarse.

81) ***Accesibilidad y diseño universal***. En el ámbito específico de las relaciones entre empresas y consumidores, los sistemas deberían estar centrados en el usuario y diseñarse de un modo que permitan que todas las personas utilicen los productos o servicios de IA con independencia de su edad, género, capacidades o características. La accesibilidad de esta tecnología para las personas con discapacidad, que están presentes en todos los grupos sociales, reviste una importancia particular. Los sistemas de IA deben ser adaptables y tener en cuenta los principios del Diseño Universal[44] para servir al mayor número posible de usuarios, observando las normas de accesibilidad pertinentes.[45] Esto permitirá un acceso equitativo y una participación activa de todas las personas en las actividades humanas informatizadas existentes y emergentes, así como en lo que atañe a las tecnologías asistenciales.[46]

82) ***Participación de las partes interesadas***. Con el fin de desarrollar sistemas de IA fiables, es recomendable consultar a las partes interesadas que se pueden ver afectadas de manera directa o indirecta por el sistema a lo largo de todo su ciclo de vida. Conviene pedir opiniones periódicamente incluso después del despliegue de los sistemas de IA y establecer mecanismos para la participación de las partes interesadas a largo plazo, por ejemplo garantizando la información, consulta y participación de los trabajadores a lo largo de todo el proceso de implantación de este tipo de sistemas en las organizaciones.

6. Bienestar social y ambiental

---

42 Puede consultarse una definición de discriminación directa e indirecta, por ejemplo, en el artículo 2 de la Directiva 2000/78/CE del Consejo, de 27 de noviembre de 2000, relativa al establecimiento de un marco general para la igualdad de trato en el empleo y la ocupación. Véase también el artículo 21 de la Carta de los Derechos Fundamentales de la UE.

43 Véase el artículo publicado por la Agencia de los Derechos Fundamentales de la Unión Europea: «BigData: Discrimination in data-supported decision making (2018)» http://fra.europa.eu/en/publication/2018/big-data-discrimination.

El artículo 42 de la Directiva sobre contratación pública exige que las especificaciones técnicas tengan en cuenta la accesibilidad y el diseño para todas las personas.

45 Por ejemplo, la norma EN 301 549.

46 Este requisito está relacionado con la Convención de las Naciones Unidas sobre los Derechos de las Personas con Discapacidad.

83) En consonancia con los *principios de equidad* y *prevención del daño*, se debería tener en cuenta también a la sociedad en su conjunto, al resto de seres sensibles y al medio ambiente como partes interesadas a lo largo de todo el ciclo de vida de la IA. Se debería fomentar la sostenibilidad y la responsabilidad ecológica de los sistemas de IA, así como impulsar la investigación de soluciones de inteligencia artificial para hacer frente a los temas que suscitan preocupación a escala mundial, como los Objetivos de Desarrollo Sostenible. Lo ideal es que la IA se utilice en beneficio de todos los seres humanos, incluidas las generaciones futuras.

84) ***Una IA sostenible y respetuosa con el medio ambiente***. Los sistemas de inteligencia artificial prometen ayudar a abordar algunas de las preocupaciones sociales más urgentes; no obstante, se debe garantizar que lo hagan del modo más respetuoso posible con el medio ambiente. En ese sentido, debería evaluarse en su integridad el proceso de desarrollo, despliegue y utilización de sistemas de IA, así como toda su cadena de suministro, a través, por ejemplo, de un examen crítico del uso de los recursos y del consumo de energía durante la formación, dando prioridad a las opciones menos perjudiciales. Se deberían promover medidas que garanticen el respeto del medio ambiente por parte de todos los eslabones de la cadena de suministro.

85) ***Impacto social***. La exposición ubicua a los sistemas sociales de IA[47] en todas las esferas de nuestra vida (sea en ámbitos como la educación, el trabajo, el cuidado o el entretenimiento) pueden alterar nuestra concepción de la acción social o afectar a nuestras relaciones y vínculos sociales. Aunque los sistemas de IA se pueden utilizar para mejorar las competencias sociales,[48] también pueden contribuir a su deterioro. Esto puede afectar al bienestar físico y mental de las personas. Por lo tanto, será necesario tener en cuenta y llevar a cabo un seguimiento exhaustivo de los efectos de esos sistemas.

86) ***Sociedad y democracia.*** Además de evaluar el impacto que ejerce el desarrollo, despliegue y utilización de un sistema de IA sobre las personas, se deberían evaluar también sus repercusiones desde la perspectiva social, teniendo en cuenta sus efectos sobre las instituciones, la democracia y la sociedad en su conjunto. El uso de sistemas de IA debería ser objeto de un estudio pormenorizado, sobre todo en situaciones relacionadas con el proceso democrático, no solo en el terreno de la adopción de decisiones políticas sino también en contextos electorales.

7. Rendición de cuentas

87) Los requisitos anteriores se complementan con el de rendición de cuentas, estrechamente relacionado con el *principio de equidad*. Este requisito exige establecer mecanismos que permitan garantizar la responsabilidad y rendición de cuentas sobre los sistemas de IA y sus resultados, tanto antes de su implantación como después de esta.

88) ***Auditabilidad.*** La auditabilidad es la capacidad para evaluar los algoritmos, los datos y los procesos de diseño. Esto no implica necesariamente que siempre deba disponerse de forma inmediata de la información sobre los modelos de negocio y la propiedad intelectual del sistema de IA. La evaluación por parte de auditores internos y externos y la disponibilidad de los correspondientes informes de evaluación pueden contribuir a la fiabilidad de esta tecnología. En aplicaciones que afecten a los derechos fundamentales, incluidas las aplicaciones esenciales desde el punto de vista de la seguridad, los sistemas de IA deberían poder someterse a auditorías independientes.

89) ***Minimización de efectos negativos y notificación de estos***. Es preciso garantizar tanto la capacidad de

[47] Esta expresión denota los sistemas de IA que se comunican e interactúan con los seres humanos mediante la simulación de la socialidad en la interacción entre humanos y robots (IA integrada) o en forma de avatares en el ámbito de la realidad virtual. De este modo, dichos sistemas tienen el potencial de transformar nuestras prácticas socioculturales y el tejido de nuestra vida social.

[48] Véase, por ejemplo, el proyecto financiado por la UE dedicado al desarrollo de software basado en IA que posibilita que los robots interactúen más eficazmente con niños autistas en sesiones de terapia dirigidas por humanos, ayudando a mejorar sus aptitudes sociales y de comunicación: http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968.

informar sobre las acciones o decisiones que contribuyen a un determinado resultado del sistema como de responder a las consecuencias de dicho resultado. La identificación, evaluación, notificación y minimización de los posibles efectos negativos de los sistemas de IA resulta especialmente crucial para quienes resulten (in)directamente afectados por ellos. Debe protegerse debidamente a los denunciantes anónimos, las ONG, los sindicatos u otras entidades que trasladen preocupaciones legítimas en relación con un sistema basado en IA. La utilización de evaluaciones de impacto (como, por ejemplo, los «equipos rojos» o determinados tipos de evaluación algorítmica de impacto) antes y después del desarrollo, despliegue y utilización de sistemas de IA puede resultar útil para minimizar sus efectos negativos. Estas evaluaciones deben ser proporcionadas al riesgo que planteen los sistemas de IA.

90) ***Búsqueda de equilibrios.*** A la hora de aplicar los requisitos anteriores pueden surgir tensiones entre ellos, por lo que puede ser necesario buscar el equilibrio. Este tipo de situaciones deberían abordarse de manera racional y metódica de acuerdo con el nivel técnico actual. Esto significa que se deberían identificar los intereses y valores subyacentes al sistema de IA y que, en el caso de que surjan conflictos, se deberá explicitar cómo se ha intentado buscar el equilibrio entre ellos y evaluar dicho equilibrio en términos del riesgo que plantea para los principios éticos, incluidos los derechos fundamentales. En las situaciones en que no sea posible identificar equilibrios aceptables desde el punto de vista ético, no se debería continuar con el desarrollo, despliegue y utilización del sistema de IA en la forma prevista. Cualquier decisión sobre la búsqueda de equilibrios debe razonarse y documentarse convenientemente. El encargado de la adopción de decisiones debe ser responsable de la forma en que se busque el equilibrio en cuestión, y revisar constantemente la idoneidad de la decisión resultante para garantizar que se puedan introducir los cambios necesarios en el sistema cuando sea preciso. [49]

91) ***Compensaciones***. Cuando se produzcan efectos adversos injustos, deberían preverse mecanismos accesibles que aseguren una compensación adecuada[50]. El hecho de saber que se podrá obtener una reparación si las cosas no salen según lo previsto es crucial para garantizar la confianza. Se debería prestar atención a las personas o grupos vulnerables.

2. Métodos técnicos y no técnicos para hacer realidad la IA fiable

92) Para cumplir los requisitos anteriormente expuestos, cabe utilizar tanto métodos técnicos como de otro tipo. Dichos métodos abarcan todas las fases del ciclo de vida de un sistema de IA. Debería llevarse a cabo una evaluación constante de los métodos empleados para cumplir los requisitos; asimismo, se debería informar y justificar[51] en todo momento los cambios introducidos en el proceso de aplicación de estos. Dado que los sistemas de IA evolucionan sin cesar y actúan en un entorno dinámico, la realización de la IA fiable es un proceso continuo, como muestra la ilustración 3.

---

49 Existen diferentes modelos de gobernanza que pueden ayudar a lograr este objetivo. Por ejemplo, la presencia de un experto o consejo ético (y específico del sector) interno o externo podría resultar útil para destacar las áreas en las que pueden surgir conflictos y sugerir la mejor forma de resolver estos. La celebración de consultas y debates pertinentes con las partes interesadas, incluidas las que corren el riesgo de verse perjudicadas por un sistema de IA, también es de gran ayuda. Las universidades europeas deberían asumir un papel de liderazgo en la formación de los especialistas en ética necesarios.

50 Véase también el dictamen de la Agencia de los Derechos Fundamentales de la Unión Europea (2017) sobre la mejora del acceso a compensaciones en el ámbito de los negocios y los derechos humanos a escala de la Unión Europea, https://fra.europa.eu/en/opinion/2017/business-human-rights.

51 Esto conlleva, por ejemplo, la justificación de las decisiones adoptadas en relación con el diseño, el desarrollo y el despliegue del sistema para incorporar los requisitos anteriormente mencionados.

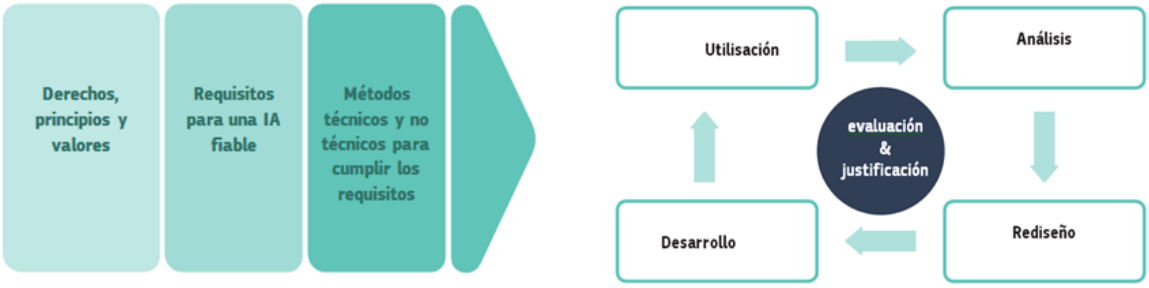

Ilustración 3: La construcción de una IA fiable a lo largo de todo el ciclo de vida del sistema

93) Los métodos siguientes pueden considerarse complementarios o alternativos entre sí, dado que los diferentes requisitos —y las diversas sensibilidades— pueden plantear la necesidad de utilizar métodos de aplicación distintos. La descripción que sigue no pretende ser exhaustiva, completa ni de obligado cumplimiento. Su finalidad es ofrecer una lista de métodos propuestos que pueden resultar útiles para lograr una IA fiable.

1. Métodos técnicos

94) En esta sección se describe una serie de métodos técnicos para garantizar la fiabilidad de la IA que se pueden incorporar en las fases de diseño, desarrollo y utilización de un sistema de IA. Los métodos que se enumeran a continuación presentan un nivel de madurez variable[52].

- *Arquitecturas para una IA fiable*

95) Los requisitos de una IA fiable deben «traducirse» en procedimientos (o en la imposición de restricciones sobre estos) que deben integrarse en la arquitectura de los sistemas de IA. Esto puede lograrse a través de un conjunto de normas de tipo «lista blanca» (comportamientos o estados) que el sistema debería seguir en todo momento, restricciones («lista negra») sobre determinados comportamientos o estados que el sistema jamás debería transgredir y combinaciones de ambas, o garantías demostrables más complejas sobre el comportamiento del sistema. El control del cumplimiento de dichas restricciones por parte del sistema durante su funcionamiento puede llevarse a cabo a través de un proceso separado.

96) Los sistemas de IA con capacidades de aprendizaje que pueden adaptar su conducta de forma dinámica pueden entenderse como un sistema no determinista que podría exhibir un comportamiento inesperado. Este tipo de sistemas se ve a menudo a través del prisma teórico de un ciclo «sentir-planear-actuar». La adaptación de esta arquitectura para garantizar la fiabilidad de la IA exige integrar los requisitos descritos en las tres etapas del ciclo: i) en la etapa «sentir», el sistema debería desarrollarse de modo que reconozca todos los elementos del entorno necesarios para garantizar el cumplimiento de los requisitos; ii) en la etapa «planear», el sistema debería tener en consideración únicamente aquellos planes que cumplan los requisitos; iii) en la etapa «actuar», las acciones del sistema deberían limitarse a comportamientos que cumplan los requisitos.

97) La arquitectura que aquí se esboza es de carácter genérico y únicamente ofrece una descripción imperfecta para la mayoría de los sistemas de IA. No obstante, también proporciona puntos de partida para las restricciones y políticas que deberían reflejarse en módulos específicos a fin de crear un sistema global fiable y que sea percibido como tal.

- *Ética y estado de Derecho desde el diseño*

[52] Pese a que algunos de ellos ya están disponibles en la actualidad, otros todavía requieren investigaciones adicionales. Aquellas áreas en las que es preciso continuar investigando también aportarán información de cara a la elaboración del segundo entregable del Grupo de expertos de alto nivel sobre inteligencia artificial, a saber, las recomendaciones sobre políticas e inversión.

98) Los métodos dirigidos a garantizar determinados valores desde el propio diseño ofrecen vínculos precisos y explícitos entre los principios abstractos que se exige que cumpla el sistema y las decisiones específicas relativas a su aplicación. La idea de que el cumplimiento de las normas puede introducirse en el diseño del sistema de IA es clave en este método. Las empresas son responsables de identificar los efectos de sus sistemas de IA desde el principio, así como las normas que deberían cumplir dichos sistemas para evitar efectos negativos. En la actualidad ya se utilizan numerosos conceptos diferentes que incluyen la expresión «desde el diseño», como la *privacidad desde el diseño* o la *seguridad desde el diseño.* Como se ha indicado anteriormente, la confianza en la IA depende de la seguridad de sus procesos, datos y resultados, así como de un diseño robusto que le permita hacer frente a posibles datos y ataques malintencionados. Los sistemas de IA deberían incluir un mecanismo de apagado a prueba de fallos, que además posibilite la reanudación del funcionamiento del sistema tras un apagado forzado (por ejemplo, un ataque).

- *Métodos de explicación*

99) Para que un sistema sea fiable, hemos de ser capaces de comprender por qué se comportó de una determinada manera y por qué ofreció una interpretación específica. Existe todo un campo de investigación, la IA explicable (conocido por las siglas XAI), que intenta resolver esta cuestión a fin de entender mejor los mecanismos subyacentes a estos sistemas y encontrar soluciones. Hoy en día este sigue siendo un desafío abierto para los sistemas de IA basados en redes neuronales. Los procesos de formación con redes neuronales pueden dar lugar a parámetros de red configurados con valores numéricos difíciles de correlacionar con resultados. Además, a veces unas pequeñas variaciones en los valores de los datos pueden traducirse en interpretaciones completamente diferentes, provocando, por ejemplo, que el sistema confunda un autobús escolar con un avestruz. Esta vulnerabilidad también se puede explotar durante los ataques contra el sistema. Los métodos que incluyen investigaciones XAI resultan vitales, no solo para explicar a los usuarios el comportamiento de los sistemas, sino también para desplegar tecnología fiable.

- *Realización de ensayos y validación*

100) Debido a la naturaleza no determinista y específica del contexto de los sistemas de IA, los ensayos tradicionales no bastan. Los errores de los conceptos y representaciones que utiliza el sistema podrían manifestarse únicamente cuando se aplique un programa a datos suficientemente realistas. En consecuencia, para verificar y validar el tratamiento de los datos, debe llevarse a cabo un seguimiento minucioso de la estabilidad, solidez y funcionamiento del modelo subyacente, tanto durante la formación como durante el despliegue, dentro de unos límites predecibles y correctamente entendidos. Debe garantizarse que el resultado del proceso de planificación sea coherente con la información introducida, y que las decisiones se adopten de un modo que permita la validación del proceso subyacente.

101) El ensayo y la validación del sistema deben tener lugar lo antes posible, garantizando que el sistema se comporte según lo previsto a lo largo de todo su ciclo de vida y, en especial, tras su despliegue. Dicha fase de ensayo y validación debería abarcar todos los componentes de un sistema de IA, incluidos los datos, los modelos previamente formados, los entornos y el comportamiento del sistema en su conjunto. Su diseño y ejecución deberían correr a cargo de un grupo de personas lo más diverso posible. Se deberían desarrollar múltiples parámetros que engloben las categorías sometidas a verificación, con el fin de obtener diferentes perspectivas. Se puede estudiar la posibilidad de realizar ensayos mediante procedimientos contradictorios a cargo de «equipos rojos» diversos y de confianza, en los que se intente «destruir» deliberadamente el sistema para encontrar posibles vulnerabilidades, así como de conceder premios a agentes externos que sean capaces de detectar e informar de forma responsable sobre los errores y debilidades del sistema. Por último, debe garantizarse que los productos o acciones sean coherentes con los resultados de los procesos precedentes, para lo cual será necesario compararlos con las políticas previamente definidas a fin de garantizar que no se vulneren.

- *Indicadores de calidad del servicio*

102) Existe la posibilidad de definir indicadores adecuados de calidad del servicio para los sistemas de IA con objeto de garantizar que sea posible saber si dichos sistemas se han ensayado y desarrollado teniendo presentes las consideraciones relativas a la seguridad. Estos indicadores podrían incluir parámetros para evaluar los ensayos realizados y la formación de algoritmos, así como los parámetros tradicionales de funcionalidad del software, su rendimiento, usabilidad, fiabilidad, seguridad y facilidad de mantenimiento.

2. Métodos no técnicos

103) En esta sección se describen diversos métodos no técnicos que pueden resultar muy útiles para garantizar y mantener la fiabilidad de la IA. Estos métodos también deberían someterse a una **evaluación constante**.

- *Normativa*

104) Como ya se ha mencionado anteriormente, hoy en día ya existe normativa de apoyo a la fiabilidad de la IA. Piénsese, por ejemplo, en la legislación sobre seguridad de los productos y en los marcos de responsabilidad. En la medida en que se considere que puede ser necesario revisar o adaptar dicha normativa —o introducir normas nuevas—, tanto con fines de protección como de fomento, dicha necesidad se planteará en nuestro segundo entregable, a saber, las recomendaciones en materia de políticas e inversión en el ámbito de la IA.

- *Códigos de conducta*

105) Las organizaciones y partes interesadas pueden adoptar estas directrices y adaptar sus cartas de responsabilidad empresarial, sus indicadores clave de rendimiento («KPI»), sus códigos de conducta o sus documentos internos de política para contribuir a los esfuerzos conducentes a la creación de una IA fiable. Desde un punto de vista más general, una organización que trabaje en un sistema de IA puede documentar sus intenciones y sustentarlas en determinados valores considerados deseables, como los derechos fundamentales, la transparencia o el principio de no causar daño.

- *Normalización*

106) Las normas, como las relativas al diseño, la fabricación o las prácticas empresariales, pueden funcionar como un sistema de gestión de la calidad para los usuarios de la IA, los consumidores, las organizaciones, las instituciones de investigación y los gobiernos, ofreciendo a todos estos agentes la capacidad de reconocer y fomentar una conducta ética a través de sus decisiones de compra. Más allá de las normas convencionales, existen enfoques de regulación conjunta: sistemas de acreditación, códigos éticos profesionales o normas dirigidas a garantizar que el diseño respete los derechos fundamentales. Entre los ejemplos existentes en la actualidad se encuentran, por ejemplo, las normas ISO o la serie de normas IEEE P7000, aunque en el futuro podría resultar adecuado crear un sello de «IA fiable» que, a partir de las normas técnicas especificadas, confirme por ejemplo que el sistema cumple los requisitos de seguridad, solidez técnica y explicabilidad.

- *Certificación*

107) Dado que no cabe esperar que todas y cada una de las personas comprendan plenamente el funcionamiento y los efectos de los sistemas de IA, puede tenerse en consideración a aquellas organizaciones que puedan acreditar ante el público que un sistema de IA es transparente, responsable y equitativo[53]. Estas certificaciones aplicarían normas desarrolladas para diferentes ámbitos de aplicación y técnicas de IA, convenientemente alineadas con las normas industriales y sociales del contexto específico de que se trate. No obstante, la certificación nunca puede sustituir a la responsabilidad. Por lo tanto, debería complementarse con marcos de rendición de cuentas que incluyan cláusulas de exención de responsabilidad, así como con mecanismos de

[53] Tal como defiende, por ejemplo, la Iniciativa para un Diseño Alineado con la Ética del Instituto de ingenieros eléctricos y electrónicos (IEEE): https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

revisión y corrección[54].

- *Rendición de cuentas a través de marcos de gobernanza*

108) Las organizaciones deberían crear marcos de gobernanza tanto internos como externos que garanticen la rendición de cuentas sobre las dimensiones éticas de las decisiones asociadas con el desarrollo, despliegue y utilización de IA. Esto puede incluir, por ejemplo, el nombramiento de una persona encargada de las cuestiones éticas relacionadas con la IA, o la creación de un panel o consejo de ética interno o externo. Entre las posibles funciones de esta persona, panel o consejo figuran la supervisión y el asesoramiento. Como se ha indicado anteriormente, las especificaciones u organismos de certificación también pueden desempeñar un papel con este fin. Se debería garantizar la existencia de canales de comunicación con grupos de supervisión públicos o industriales, el intercambio de buenas prácticas, el debate sobre dilemas o la notificación de problemas emergentes que susciten preocupación desde el punto de vista ético. Estos mecanismos pueden complementar —pero no sustituir— la supervisión legal (por ejemplo, en forma de designación de un funcionario responsable de la protección de datos u otras medidas equivalentes, según requieran las leyes de protección de datos).

- *Educación y concienciación para fomentar una mentalidad ética*

109) La IA fiable promueve la participación con conocimiento de causa de todas las partes interesadas. La comunicación, la educación y la formación desempeñan un papel importante, tanto para garantizar la difusión del conocimiento sobre los efectos potenciales de los sistemas de IA como para que la ciudadanía tome conciencia de que puede participar en la configuración del desarrollo social. Esto incluye a todas las partes interesadas, por ejemplo las implicadas en la fabricación de productos (los diseñadores y desarrolladores), los usuarios (empresas o individuos) y otros grupos afectados (que quizá no adquieran o utilicen un sistema de IA, pero a quienes afectan las decisiones de estos sistemas, así como la sociedad en su conjunto). Se debería impulsar la educación básica sobre la IA entre la sociedad para garantizar que la población cuente con los conocimientos esenciales sobre ella. Un requisito previo para ello es garantizar unas adecuadas competencias y formación de los expertos en ética en este espacio.

- *Participación de las partes interesadas y diálogo social*

110) La IA ofrece numerosos beneficios y Europa necesita garantizar que toda la población pueda disfrutar de ellos. Esto requiere un debate abierto y la implicación de los interlocutores sociales, las partes interesadas y el público en general. Muchas organizaciones recurren ya a paneles de partes interesadas para debatir sobre el uso de sistemas de IA y análisis de datos. En estos paneles participan diferentes tipos de personas, como expertos jurídicos o técnicos, especialistas en ética, representantes de los consumidores o trabajadores. El fomento de la participación y el diálogo sobre la utilización y los efectos de los sistemas de IA respalda la evaluación de resultados y enfoques, y puede resultar particularmente útil en casos complejos.

- *Diversidad y equipos de diseño inclusivos*

111) La diversidad y la inclusión desempeñan un papel esencial al desarrollar los sistemas de IA que se utilizarán en el mundo real. Es crucial que, a medida que los sistemas de IA vayan desempeñando una mayor cantidad de tareas por sí mismos, los equipos encargados de la adquisición o del diseño, desarrollo, ensayo, mantenimiento y despliegue de estos sistemas reflejen la diversidad de los usuarios y de la sociedad en general. Esto contribuye a garantizar la objetividad y la toma en consideración de las diferentes perspectivas, necesidades y objetivos. Lo ideal es que la diversidad no solo se materialice en los equipos en términos de género, cultura y edad, sino también de antecedentes profesionales y conjuntos de competencias.

[54] Para obtener más información sobre las limitaciones de la certificación, véase: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

Orientaciones clave derivadas del capítulo II:

- ✓ Garantizar que los sistemas de IA satisfagan, a lo largo de todo su ciclo de vida, los requisitos para una IA fiable: 1) acción y supervisión humanas, 2) solidez técnica y seguridad, 3) gestión de la privacidad y de los datos, 4) transparencia, 5) diversidad, no discriminación y equidad, 6) bienestar ambiental y social, y 7) rendición de cuentas.
- ✓ Para garantizar el cumplimiento de estos requisitos, se deberá estudiar la posibilidad de emplear tanto métodos técnicos como no técnicos.
- ✓ Impulsar la investigación y la innovación para ayudar a evaluar los sistemas de IA y a promover el cumplimiento de los requisitos; divulgar los resultados y las preguntas de interpretación abierta al público en general, y formar sistemáticamente a una nueva generación de especialistas en ética de la IA.
- ✓ Comunicar información a las partes interesadas, de un modo claro y proactivo, sobre las capacidades y limitaciones de los sistemas de IA, posibilitando el establecimiento de expectativas realistas, así como sobre el modo en que se cumplen los requisitos. Ser transparentes acerca del hecho de que se está trabajando con un sistema de IA.
- ✓ Facilitar la trazabilidad y la auditabilidad de los sistemas de IA, especialmente en contextos y situaciones críticos.
- ✓ Implicar a las partes interesadas en todo el ciclo de vida de los sistemas de IA. Promover la formación y la educación, de manera que todas las partes interesadas sean conocedoras de la IA fiable y reciban formación en la materia.
- ✓ Ser conscientes de que pueden existir tensiones fundamentales entre los diferentes principios y requisitos. Identificar, evaluar, documentar y comunicar constantemente este tipo de tensiones y sus soluciones.

III. Capítulo III: Evaluación de la IA fiable

112) Con base en los requisitos clave definidos en el capítulo II, se ofrece en este capítulo una **lista no exhaustiva para la evaluación de la fiabilidad de la IA** (versión piloto) con el fin de **poner en práctica la IA fiable**. La lista es de aplicación, en particular, a los sistemas de IA que interactúen directamente con los usuarios, y va dirigida fundamentalmente a desarrolladores y responsables del despliegue de sistemas de IA (sean desarrollados internamente o adquiridos a terceros). La lista de evaluación no aborda la puesta en práctica del primer componente de la IA fiable (la IA lícita). La utilización de esta lista no constituye una prueba del cumplimiento legal ni pretende servir como guía para garantizar el cumplimiento de la legislación vigente. Dado el carácter específico de las aplicaciones de los sistemas de IA, será necesario adaptar la lista de evaluación a los casos de uso y contextos específicos en los que operen dichos sistemas. Además, este capítulo ofrece una recomendación general acerca de cómo aplicar la lista de evaluación para una IA fiable a través de una estructura de gobernanza que abarque tanto el nivel operativo como el de gestión.

113) La lista de evaluación y la estructura de gobernanza se desarrollarán en estrecha colaboración con las partes interesadas tanto del sector público como del privado. El proceso tendrá carácter piloto y posibilitará la obtención de numerosos comentarios, observaciones y opiniones mediante dos procesos paralelos:

a. un proceso cualitativo que garantizará la representatividad, en el que un reducido número de empresas, organizaciones e instituciones seleccionadas (de diferentes tamaños y pertenecientes a distintos sectores) se inscribirán para experimentar la lista de evaluación y la estructura de gobernanza en la práctica, tras lo cual ofrecerán sus comentarios detallados al respecto;

b. un proceso cuantitativo, en el que todas las partes interesadas podrán inscribirse para

experimentar la lista de evaluación y dar su opinión a través de una consulta abierta.

114) Tras la fase piloto, los resultados del proceso de formulación de observaciones se integrarán en la lista de evaluación y se elaborará una versión revisada de esta a principios de 2020. El objetivo es desarrollar un marco que se pueda utilizar de forma horizontal en todas las aplicaciones, por lo que servirá de base para garantizar la fiabilidad de la IA en todos los ámbitos. Una vez establecida dicha base, se podría desarrollar un marco sectorial o para aplicaciones específicas.

Gobernanza

115) Las empresas, organizaciones e instituciones pueden tener interés en estudiar de qué modo se puede aplicar la lista de evaluación de la IA fiable en sus respectivas entidades. Esto se puede llevar a cabo incluyendo el proceso de evaluación en los mecanismos de gobernanza existentes o bien introduciendo procesos nuevos. La elección dependerá de la estructura interna de la organización, de su tamaño y de los recursos disponibles.

116) Las investigaciones disponibles[55] demuestran que, para lograr el cambio, es esencial contar con la atención de la dirección al más alto nivel. Asimismo, dichos estudios ponen de manifiesto que el hecho de involucrar a todas las partes interesadas de una empresa, organización o institución favorece la aceptación y la pertinencia de la introducción de cualquier nuevo proceso (sea de naturaleza tecnológica o no)[56]. Por lo tanto, recomendamos implicar en el proceso tanto al nivel operativo como a la dirección superior de la organización.

Nivel	**Funciones pertinentes (dependiendo de la organización)**
Dirección y consejo de administración	La dirección superior debate sobre el desarrollo, despliegue o adquisición de la IA y evalúa dichos aspectos, y actúa a modo de órgano superior para evaluar todas las innovaciones y usos de la IA cuando se detecten preocupaciones cruciales. Implica a los afectados por la posible introducción de sistemas de IA (como, por ejemplo, los trabajadores) y sus representantes a lo largo de todo el proceso a través de procedimientos de información, consulta y participación.
Departamento de cumplimiento normativo/jurídico/de responsabilidad empresarial	El departamento de responsabilidad supervisa el uso de la lista de evaluación y su necesaria evolución para adaptarla a los cambios tecnológicos y reglamentarios. Actualiza las normas o las políticas internas aplicables a los sistemas de IA y garantiza que la utilización de estos se ajuste al marco jurídico y reglamentario vigente y a los valores de la organización.
Departamento de desarrollo de productos y servicios o equivalente	El departamento de desarrollo de productos y servicios utiliza la lista de evaluación para evaluar los productos y servicios basados en IA y documenta todos los resultados. Estos resultados se debaten en la dirección, que es la responsable última de aprobar las aplicaciones basadas en IA nuevas o revisadas.
Garantía de calidad	El departamento de garantía de calidad (o equivalente) garantiza y comprueba los resultados de la lista de evaluación y adopta las medidas oportunas para trasladar un problema a una instancia superior de decisión en el caso de que el resultado no sea

[55] https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation.

[56] Véase, por ejemplo, A. Bryson, E. Barth y H. Dale-Olsen, «The Effects of Organisational change on worker well-being and the moderating role of trade unions», *ILRReview*, 66(4), julio de 2013; Jirjahn, U. y Smith, S.C. (2006). «What Factors Lead Management to Support or Oppose Employee Participation—With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations», 45(4), 650–680; Michie, J. y Sheehan, M. (2003). «Labour market deregulation, "flexibility" and innovation», *Cambridge Journal of Economics*, 27(1), 123–143.

	satisfactorio o que se detecten resultados imprevistos.
Recursos humanos	El departamento de RR. HH. garantiza la combinación adecuada de competencias y la diversidad de perfiles de los desarrolladores de sistemas de IA. Se asegura de que se proporcione el nivel de formación adecuado sobre la IA fiable en el seno de la organización.
Adquisiciones	El departamento de compras o adquisiciones se cerciora de que el proceso de adquisición de productos o servicios basados en IA incluya una verificación de la fiabilidad de la IA.
Operaciones cotidianas	Los desarrolladores y los directores de proyectos incluyen la lista de evaluación en su trabajo diario y documentan los resultados de la evaluación.

Cómo utilizar la lista de evaluación para una IA fiable

117) Cuando se utilice la lista de evaluación en la práctica, recomendamos que no se preste atención únicamente a aquellas áreas que susciten preocupación, sino también a las preguntas que no sea posible responder (fácilmente). Un posible problema puede ser la falta de diversidad de las cualificaciones y competencias del equipo encargado del desarrollo y ensayo del sistema de IA, por lo que podría ser necesario involucrar a otras partes interesadas de dentro o fuera de la organización. Se recomienda vivamente registrar todos los resultados tanto en términos técnicos como de gestión, asegurando que la resolución de problemas se entienda correctamente en todos los niveles de la estructura de gestión.

118) La presente lista de evaluación tiene la finalidad de guiar a los profesionales de la IA en el desarrollo, despliegue y utilización de una IA fiable. La evaluación deberá adaptarse de manera proporcionada al caso de uso específico de que se trate. Durante la fase piloto podrían aparecer determinadas áreas sensibles; en esos casos, se evaluará en el paso siguiente la necesidad de introducir especificaciones adicionales. Pese a que esta lista de evaluación no ofrece respuestas concretas a las preguntas planteadas, anima a reflexionar sobre los pasos que pueden ayudar a garantizar la fiabilidad de los sistemas de IA y sobre las medidas que deberían adoptarse en ese sentido.

Relación con leyes y procesos existentes

119) También es importante para las personas involucradas en el desarrollo, despliegue y utilización de IA saber que existen varias leyes que prescriben determinados procesos y prohíben ciertos resultados; estas leyes pueden solaparse y coincidir con algunas de las medidas enumeradas en la lista de evaluación. A modo de ejemplo, la legislación en materia de protección de datos establece una serie de requisitos legales que deben satisfacer las personas implicadas en la recopilación y el tratamiento de datos personales. No obstante, dado que la IA fiable también exige guiarse por criterios éticos en el manejo de los datos, la existencia de procedimientos y políticas internos dirigidos a garantizar el cumplimiento de las leyes de protección de datos también podría contribuir a facilitar el manejo ético de los datos y, por tanto, complementar los procesos legales existentes. La utilización de esta lista *no* constituye, sin embargo, una prueba del cumplimiento legal ni pretende servir como guía para garantizar el cumplimiento de la legislación vigente. En lugar de ello, el objetivo de esta lista de evaluación es ofrecer un conjunto de preguntas concretas dirigidas a aquellos destinatarios que estén intentando asegurarse de que su enfoque de desarrollo y despliegue de IA esté orientado a crear —y trate de asegurar— una IA fiable.

120) De forma similar, muchos profesionales de la IA cuentan ya con herramientas de evaluación y procesos de desarrollo de software para garantizar el cumplimiento de normas no jurídicas. La evaluación que sigue no debería llevarse a cabo como un ejercicio autónomo, sino integrarse en dichas prácticas.

LISTA DE EVALUACIÓN PARA UNA IA FIABLE (VERSIÓN PILOTO)

1. Acción y supervisión humanas

Derechos fundamentales:

- ✓ En aquellos casos de usos en los que puedan producirse efectos potencialmente negativos para los derechos fundamentales, ¿ha llevado usted a cabo una evaluación del impacto sobre los derechos fundamentales? ¿Ha identificado y documentado los posibles equilibrios entre los diferentes principios y derechos?

- ✓ ¿Interactúa el sistema de IA con el proceso de adopción de decisiones por parte de usuarios finales humanos (por ejemplo, con las acciones recomendadas, las decisiones que es preciso adoptar o la presentación de opciones)?

  - ¿Existe en esos casos el riesgo de que el sistema de IA afecte a la autonomía humana al interferir con el proceso de adopción de decisiones del usuario final de forma imprevista?

  - ¿Ha considerado usted si el sistema de IA debería informar a los usuarios de que una decisión, contenido, recomendación o resultado es fruto de una decisión algorítmica?

  - En el caso de que el sistema de IA cuente con un bot de charla o un sistema conversacional, ¿son los usuarios finales humanos conocedores de que están interactuando con un agente no humano?

Acción humana:

- ✓ En el caso de que el sistema de IA se implante en el proceso de trabajo, ¿ha tenido usted en cuenta la asignación de tareas entre el sistema de IA y los trabajadores humanos para garantizar interacciones adecuadas y una supervisión y control humanas apropiadas?

  - ¿El sistema de IA mejora o aumenta las capacidades humanas?

  - ¿Se han adoptado medidas para evitar que los procesos de trabajo confíen o dependan en exceso del sistema de IA?

Supervisión humana:

- ✓ ¿Ha analizado cuál sería el nivel adecuado de control humano sobre el sistema de IA específico y para el caso de uso concreto de que se trate?

  - ¿Puede describir el nivel de control o implicación humana, si procede? ¿Quién es la persona que ostenta el control del sistema y cuáles son los momentos o herramientas para la intervención humana?

  - ¿Ha establecido mecanismos y adoptado medidas para garantizar la posibilidad de dicho control o supervisión humanos o para asegurar que las decisiones se tomen bajo la responsabilidad exclusiva de seres humanos?

  - ¿Ha adoptado alguna medida para posibilitar la realización de auditorías y para solucionar cualquier problema relacionado con la gestión de la autonomía de la IA?

- ✓ En el caso de que exista un sistema de IA (o un caso de uso) autónomo o con capacidad de autoaprendizaje, ¿ha establecido mecanismos de control y supervisión más concretos?
  - ¿Qué tipo de mecanismos de detección y respuesta ha establecido para evaluar si algo puede salir mal?
  - ¿Se ha asegurado de disponer de un botón de desconexión o un procedimiento que permita abortar una operación en condiciones de seguridad en caso necesario? ¿Implica ese procedimiento que se aborta el proceso en su totalidad, en parte o la delegación del control a un ser humano?

2. <u>Solidez técnica y seguridad</u>

Resistencia a los ataques y seguridad:

- ✓ ¿Ha evaluado las posibles formas de ataque a las que puede ser vulnerable el sistema de IA?
  - En particular, ¿ha analizado los diferentes tipos y naturalezas de las vulnerabilidades, como la contaminación de los datos, la infraestructura física o los ciberataques?
- ✓ ¿Ha adoptado medidas o sistemas para garantizar la integridad del sistema de IA y su capacidad para resistir posibles ataques?
- ✓ ¿Ha evaluado el comportamiento de su sistema en situaciones o entornos imprevistos?
- ✓ ¿Ha analizado si su sistema se puede utilizar (y, en caso afirmativo, en qué medida) para diferentes fines? Si es así, ¿ha adoptado medidas adecuadas para prevenir su uso con fines no deseados (como, por ejemplo, la no divulgación de la investigación o despliegue del sistema)?

Plan de repliegue y seguridad general:

- ✓ ¿Se ha asegurado de que su sistema cuente con un plan de repliegue suficiente en el caso de que se enfrente a algún ataque malintencionado o a otro tipo de situación inesperada (por ejemplo, procedimientos técnicos de conmutación o formulación de preguntas a un ser humano antes de continuar)?
- ✓ ¿Ha analizado el nivel de riesgo que plantea el sistema de IA en el caso de uso concreto previsto?
  - ¿Ha introducido algún proceso para medir y evaluar los riesgos y la seguridad?
  - ¿Ha proporcionado la información necesaria en caso de que exista algún riesgo para la integridad física de las personas?
  - ¿Ha estudiado la posibilidad de contratar una póliza de seguro para hacer frente a los posibles daños que provoque el sistema de IA?
  - ¿Ha identificado los riesgos potenciales para la seguridad asociados a (otros) usos previsibles de la tecnología, incluidos los usos accidentales o malintencionados? ¿Existe algún plan para mitigar o gestionar esos riesgos?
- ✓ ¿Ha evaluado si es probable que el sistema de IA cause daños a los usuarios o a terceros? En caso

afirmativo, ¿ha evaluado la probabilidad, el daño potencial, el público afectado y la gravedad de tales daños?

- Si existe el riesgo de que el sistema de IA ocasione daños, ¿ha tenido en cuenta las leyes de responsabilidad civil y de protección de los consumidores? ¿Cómo?
- ¿Ha analizado los efectos potenciales o el riesgo para la seguridad del medio ambiente o de la fauna?
- ¿Ha tenido en cuenta en su análisis de riesgos si los problemas de seguridad o de la red (por ejemplo, los peligros para la ciberseguridad) plantean riesgos para la seguridad o pueden causar daños debido a un comportamiento imprevisto del sistema de IA?

✓ ¿Ha estimado el efecto probable de un fallo de su sistema de IA que provoque que el sistema ofrezca resultados erróneos, quede fuera de servicio o proporcione resultados socialmente inaceptables (como, por ejemplo, prácticas discriminatorias)?

- ¿Ha definido umbrales y mecanismos de gestión para los escenarios anteriores a fin de activar planes alternativos o de repliegue?
- ¿Ha definido y ensayado planes de repliegue?

Precisión

✓ ¿Ha evaluado qué nivel y definición de precisión se requerirá en el contexto del sistema de IA y para el caso de uso previsto?

- ¿Ha evaluado cómo se mide y garantiza la precisión?
- ¿Ha adoptado medidas para garantizar que los datos utilizados sean exhaustivos y estén actualizados?
- ¿Ha adoptado medidas para evaluar si es necesario disponer de datos adicionales, por ejemplo para mejorar la precisión o eliminar sesgos?

✓ ¿Ha evaluado los daños que se ocasionarían si el sistema de IA realizara predicciones incorrectas?

✓ ¿Ha establecido algún mecanismo para medir si el sistema está realizando una cantidad inaceptable de predicciones erróneas?

✓ Si el sistema está realizando predicciones erróneas, ¿ha establecido una serie de pasos que permitan subsanar el problema?

Fiabilidad y reproducibilidad:

✓ ¿Ha diseñado una estrategia para supervisar y verificar que el sistema cumple los objetivos, el propósito y las aplicaciones previstas?

- ¿Ha comprobado si es necesario tener en cuenta algún contexto o condición particular para garantizar la reproducibilidad?
- ¿Ha introducido procesos o métodos de verificación para medir y garantizar los diferentes aspectos de la fiabilidad y la reproducibilidad?
- ¿Ha establecido algún proceso para describir las situaciones en las que un sistema de IA falla en

determinados tipos de entornos?

- ¿Ha documentado y detallado claramente esos procesos para la verificación de la fiabilidad de los sistemas de IA?

  ¿Ha establecido algún mecanismo o comunicación para garantizar a los usuarios (finales) que el sistema de IA es fiable?

3. <u>Gestión de la privacidad y de los datos</u>

Respeto de la privacidad y de la protección de datos:

- ✓ Dependiendo del caso de uso, ¿ha establecido un mecanismo que permita notificar los problemas relacionados con la privacidad o la protección de datos en los procesos de recopilación de datos de los sistemas de IA (tanto con fines de formación como de funcionamiento) y su tratamiento?
- ✓ ¿Ha evaluado el tipo y alcance de los datos incluidos en sus bases de datos (por ejemplo, si estas contienen datos de carácter personal)?
- ✓ ¿Ha analizado formas de desarrollar el sistema de IA o de formar el modelo en las que no sea necesario utilizar datos personales o potencialmente sensibles (o que utilicen la mínima cantidad posible de este tipo de datos)?
- ✓ ¿Ha introducido mecanismos de aviso y control sobre los datos personales en función del caso de uso (como, por ejemplo, el consentimiento válido y la posibilidad de revocar el uso de dichos datos, cuando proceda)?
- ✓ ¿Ha tomado medidas para mejorar la privacidad, por ejemplo a través de procesos como el encriptado, la anonimización y la agregación?
- ✓ En los casos en que exista una persona responsable de la privacidad de los datos, ¿la ha implicado desde una fase inicial del proceso?

Calidad e integridad de los datos:

- ✓ ¿Ha alineado su sistema con las normas potencialmente pertinentes (por ejemplo, ISO, IEEE) o ha adoptado protocolos generales para la gestión y gobernanza cotidianas de sus datos?
- ✓ ¿Ha establecido mecanismos de supervisión para la recopilación, almacenamiento, tratamiento y utilización de los datos?
- ✓ ¿Ha evaluado su grado de control sobre la calidad de las fuentes de datos externas utilizadas?
- ✓ ¿Ha instaurado procesos para garantizar la calidad y la integridad de sus datos? ¿Ha estudiado la posibilidad de introducir otros procesos? ¿Cómo está verificando que sus conjuntos de datos no son vulnerados ni objeto de ataques?

Acceso a los datos:

- ✓ ¿Qué protocolos, procesos y procedimientos se han seguido para gestionar y garantizar una adecuada gobernanza de los datos?

  - ¿Ha evaluado quién puede acceder a los datos de los usuarios y en qué circunstancias?

- ¿Se ha asegurado de que esas personas poseen la cualificación para acceder a los datos, que se les exige acceder a ellos y que cuentan con las competencias necesarias para comprender los detalles de la política de protección de datos?

- ¿Ha asegurado la existencia de un mecanismo de supervisión que permita registrar cuándo, dónde, cómo y quién accede a los datos, y con qué propósito?

4. <u>Transparencia</u>

Trazabilidad:

✓ ¿Ha adoptado medidas que puedan garantizar la trazabilidad? Esto puede conllevar la documentación de:

- los métodos utilizados para diseñar y desarrollar el sistema algorítmico:
  - en el caso de un sistema de IA basado en reglas, se debería documentar el método de programación o la forma en que se creó el modelo;
  - en el caso de un sistema de IA basado en el aprendizaje, se debería documentar el método de formación del algoritmo, incluidos los datos de entrada que se recopilaron y seleccionaron y la forma en que se hizo;
- los métodos empleados para ensayar y validar el sistema algorítmico:
  - en el caso de un sistema de IA basado en reglas, se deberían documentar los escenarios o casos de uso utilizados para los ensayos y la validación;
  - en el caso de un modelo basado en el aprendizaje, se debería documentar la información sobre los datos utilizados para los ensayos y la validación;
- los resultados del sistema algorítmico:
  - se deberían documentar los resultados del algoritmo o las decisiones adoptadas por este, así como otras posibles decisiones que se producirían en casos diferentes (por ejemplo, para otros subgrupos de usuarios).

Explicabilidad:

✓ ¿Ha evaluado en qué medida son comprensibles las decisiones y, por tanto, el resultado producido por el sistema de IA?

✓ ¿Se ha asegurado de que se pueda elaborar una explicación comprensible para todos los usuarios que puedan desearla sobre las razones por las que un sistema adoptó una decisión determinada que diera lugar a un resultado específico?

✓ ¿Ha evaluado en qué medida la decisión del sistema influye en los procesos de adopción de decisiones de la organización?

✓ ¿Ha evaluado por qué se desplegó ese sistema en particular en esa área concreta?

✓ ¿Ha evaluado el modelo de negocio del sistema (por ejemplo, de qué modo crea valor para la organización)?

✓ ¿Ha diseñado el sistema de IA teniendo en mente desde el principio la interpretabilidad?

- ¿Ha investigado y tratado de utilizar el modelo más sencillo e interpretable posible para la aplicación en cuestión?
- ¿Ha evaluado si puede analizar sus datos relativos a la formación y los ensayos realizados? ¿Puede modificar y actualizar estos datos a lo largo del tiempo?
- ¿Ha evaluado si, tras la formación y el desarrollo del modelo, tiene alguna posibilidad de examinar su interpretabilidad o si dispone de acceso al flujo de trabajo interno del modelo?

Comunicación:

✓ ¿Ha informado a los usuarios (finales) —mediante cláusulas de exención de responsabilidad u otros medios— de que están interactuando con un sistema de IA y no con otro ser humano? ¿Ha etiquetado su sistema de IA como tal?

✓ ¿Ha establecido mecanismos para informar a los usuarios de las razones y criterios subyacentes a los resultados del sistema de IA?

- ¿Se han comunicado claramente estos a los usuarios previstos?
- ¿Ha establecido procesos que tengan en cuenta las opiniones de los usuarios y que utilicen dichas opiniones para adaptar el sistema?
- ¿Ha informado sobre los riesgos potenciales o percibidos, como la posible existencia de sesgos?
- ¿Ha tenido también en cuenta, según el caso de uso, la comunicación y la transparencia hacia otras audiencias, hacia terceros o hacia el público en general?

✓ ¿Ha dejado claro el propósito del sistema de IA y quién o qué podrá beneficiarse del producto o servicio que ofrezca este?

- ¿Se han especificado y se ha informado claramente sobre los escenarios de utilización del producto, estudiando posibles métodos de comunicación alternativos para garantizar que dicha información sea comprensible y adecuada para los usuarios a los que se dirige?
- Según el caso de uso, ¿ha tenido en cuenta la psicología humana y sus posibles limitaciones, como el riesgo de confusión, el sesgo de confirmación o la fatiga cognitiva?

✓ ¿Ha comunicado con claridad las características, limitaciones y posibles carencias del sistema de IA:

- en caso de desarrollo: a las personas encargadas de su despliegue en un producto o servicio?
- en caso de despliegue: a los usuarios finales o consumidores?

5. <u>Diversidad, no discriminación y equidad</u>

Necesidad de evitar sesgos injustos:

✓ ¿Se ha asegurado de que exista una estrategia o un conjunto de procedimientos para evitar crear o reforzar un sesgo injusto en el sistema de IA, tanto en relación con el uso de los datos de entrada como en lo referente al diseño del algoritmo?

- ¿Ha evaluado y reconocido las posibles limitaciones que emanan de la composición de los

conjuntos de datos utilizados?

- ¿Ha tenido en cuenta la diversidad y representatividad de los usuarios en los datos? ¿Ha realizado ensayos para poblaciones específicas o casos de uso problemáticos?
- ¿Ha investigado y utilizado las herramientas técnicas disponibles para mejorar su comprensión de los datos, el modelo y su rendimiento?
- ¿Ha establecido procesos para verificar la existencia de posibles sesgos y llevar a cabo un seguimiento de estos durante las fases de desarrollo, despliegue y utilización del sistema?

✓ Dependiendo del caso de uso, ¿se ha asegurado de introducir un mecanismo que permita a otras personas informar sobre posibles problemas relacionados con la existencia de sesgos, discriminación o un rendimiento deficiente del sistema de IA?

- ¿Ha estudiado vías y métodos de comunicación claros sobre cómo y a quién informar sobre este tipo de problemas?
- ¿Ha tenido en cuenta no solo a los usuarios (finales) sino también a otras personas que puedan verse indirectamente afectadas por el sistema de IA?

✓ ¿Ha evaluado si existe la posibilidad de que las decisiones varíen aunque las condiciones no cambien?

- Si es así, ¿ha estudiado cuáles podrían ser las causas de ello?
- En caso de variabilidad, ¿ha establecido algún mecanismo de medición o evaluación del impacto potencial de dicha variabilidad sobre los derechos fundamentales?

✓ ¿Se ha asegurado de utilizar una definición operativa adecuada de «equidad» para aplicarla en el diseño de sistemas de IA?

- ¿Se trata de una definición de uso común? ¿Estudió otras definiciones antes de optar por la seleccionada?
- ¿Ha instaurado análisis o parámetros cuantitativos para medir y poner a prueba la definición de equidad aplicada?
- ¿Ha establecido mecanismos para garantizar la equidad en sus sistemas de IA? ¿Ha considerado otros posibles mecanismos?

Accesibilidad y diseño universal:

✓ ¿Se ha asegurado de que el sistema de IA se adapte a una amplia variedad de preferencias y capacidades individuales?

- ¿Ha evaluado si las personas con discapacidad, con necesidades especiales o en riesgo de exclusión pueden utilizar el sistema de IA? ¿Cómo se integró este aspecto en el sistema y cómo se verifica su funcionamiento?
- ¿Se ha asegurado de que la información sobre el sistema de IA también sea accesible para los usuarios de tecnologías asistenciales?
- ¿Implicó o consultó a esta comunidad durante la fase de desarrollo del sistema de IA?

- ✓ ¿Ha tenido en cuenta el impacto de su sistema de IA en sus usuarios potenciales?
  - ¿Es el equipo involucrado en el desarrollo del sistema de IA representativo de la audiencia a la que va dirigido? ¿Es representativo de la población en general y tiene también en cuenta a otros grupos que pudieran verse afectados de manera tangencial por el sistema?
  - ¿Ha evaluado la posibilidad de que haya personas o grupos que puedan verse afectados de forma desproporcionada por las implicaciones negativas del sistema?
  - ¿Ha recabado la opinión de otros equipos o grupos representativos de diferentes contextos y experiencias?

Participación de las partes interesadas:

- ✓ ¿Ha estudiado la posibilidad de introducir algún mecanismo para incorporar la participación de diferentes partes interesadas en el desarrollo y la utilización del sistema de IA?
- ✓ ¿Ha allanado el camino para la introducción del sistema de IA en su organización, informando e implicando previamente a los trabajadores afectados y sus representantes?

6. <u>Bienestar social y ambiental</u>

Una IA sostenible y respetuosa con el medio ambiente:

- ✓ ¿Ha establecido mecanismos para medir el impacto ambiental del desarrollo, despliegue y utilización del sistema de IA (por ejemplo, energía consumida por cada centro de datos, tipo de energía utilizada por los centros de datos, etc.)?
- ✓ ¿Se ha asegurado de introducir medidas para reducir el impacto ambiental de su sistema de IA a lo largo de todo su ciclo de vida?

Impacto social:

- ✓ En el caso de que el sistema de IA interactúe directamente con seres humanos:
  - ¿Ha evaluado si el sistema de IA alienta a los humanos a establecer un vínculo y desarrollar la empatía con el sistema?
  - ¿Se ha asegurado de que el sistema indique claramente que su interacción social es simulada y que no tiene capacidad para «entender» ni «sentir»?
- ✓ ¿Se ha asegurado de que se entiendan correctamente los efectos sociales del sistema de IA? Por ejemplo, ¿ha evaluado si existe un riesgo de pérdida de puestos de trabajo o de descualificación de la mano de obra? ¿Qué pasos se han dado para contrarrestar esos riesgos?

Sociedad y democracia:

- ✓ ¿Ha evaluado el impacto social global asociado al uso del sistema de IA más allá del que tenga sobre el usuario (final), como, por ejemplo, las partes interesadas que pueden verse indirectamente afectadas por dicho sistema?

7. <u>Rendición de cuentas</u>

Auditabilidad:

- ✓ ¿Ha establecido mecanismos para facilitar la auditabilidad del sistema por parte de agentes internos o independientes (garantizando, por ejemplo, la trazabilidad y registro de los procesos y resultados del sistema de IA)?

Minimización de efectos negativos y notificación de estos:

- ✓ ¿Ha llevado a cabo una evaluación de riesgos o de impacto del sistema de IA que tenga en cuenta a las diferentes partes interesadas que se vean afectadas por este de forma directa o indirecta?
- ✓ ¿Ha establecido marcos de formación y educación para el desarrollo de prácticas de rendición de cuentas?
  - ¿Qué trabajadores o partes del equipo están implicados en ello? ¿Trasciende la fase de desarrollo?
  - ¿Se explica también en esa formación el posible marco jurídico aplicable al sistema de IA?
  - ¿Ha considerado la posibilidad de crear una «junta de revisión ética de la IA» u otro mecanismo similar para debatir sobre las prácticas éticas y de rendición de cuentas en general, incluidas las posibles «zonas grises»?
- ✓ Además de las iniciativas o marcos internos para supervisar la ética y la rendición de cuentas, ¿se cuenta con algún tipo de orientación externa o se han establecido también procesos de auditoría?
- ✓ ¿Existe algún proceso para que los trabajadores o agentes externos (por ejemplo, proveedores, consumidores, distribuidores/vendedores) informen sobre posibles vulnerabilidades, riesgos o sesgos en el sistema de IA o su aplicación?

Documentación de los equilibrios alcanzados:

- ✓ ¿Se ha establecido algún mecanismo para identificar los intereses y valores que implica el sistema de IA y los posibles equilibrios entre ellos?
- ✓ ¿Qué procesos ha seguido para decidir sobre los equilibrios necesarios? ¿Se ha asegurado de documentar la decisión sobre la búsqueda de dichos equilibrios?

Capacidad de obtener compensación:

- ✓ ¿Ha establecido un conjunto de mecanismos adecuado que permita obtener compensación en el caso de que se produzca cualquier daño o efecto adverso?
- ✓ ¿Se han instaurado mecanismos para proporcionar información a usuarios (finales) y a terceros sobre las oportunidades de obtener compensación?

Invitamos a todas las partes interesadas a experimentar esta lista de evaluación con carácter piloto y a hacernos llegar sus comentarios sobre su aplicabilidad, exhaustividad, pertinencia para el ámbito o aplicación específica de la IA, así como sobre la existencia de posibles solapamientos o aspectos complementarios con los procesos de cumplimiento o de evaluación existentes. A partir de los comentarios recibidos, se elaborará una versión revisada de la lista de evaluación para una IA fiable que se presentará a la Comisión a principios de 2020.

Orientaciones clave derivadas del capítulo III:

- ✓ Adoptar una **lista de evaluación** de la fiabilidad de la IA al desarrollar, desplegar o utilizar IA, y adaptarla al caso de uso específico en el que se utilice el sistema.
- ✓ Tener presente que este tipo de listas de evaluación **nunca pueden ser exhaustivas**. Garantizar la fiabilidad de la IA no consiste en marcar casillas de verificación, sino en identificar constantemente requisitos, evaluar soluciones y asegurar mejores resultados a lo largo de todo el ciclo de vida del sistema de IA, implicando a las partes interesadas en el proceso.

C. EJEMPLOS DE OPORTUNIDADES Y PREOCUPACIONES FUNDAMENTALES QUE PLANTEA LA IA

121) En la sección que sigue se ofrecen ejemplos de desarrollo y utilización de la IA que se deberían fomentar, así como ejemplos de situaciones en las que el desarrollo, despliegue o utilización de la IA pueden contravenir nuestros valores y plantear preocupaciones específicas. Es preciso encontrar un equilibrio entre lo que se puede hacer con la IA y lo que se debería hacer con ella, además de prestar la debida atención a lo que no se debería hacer con esta tecnología.

1. Ejemplos de oportunidades que ofrece una IA fiable

122) Una IA fiable puede representar una gran oportunidad para ayudar a mitigar los importantes desafíos a los que se enfrenta la sociedad, como el envejecimiento de la población, la creciente desigualdad social y la contaminación ambiental. Este potencial también tiene su reflejo a escala mundial, como, por ejemplo, en los Objetivos de Desarrollo Sostenible de las Naciones Unidas[57]. En la sección siguiente se analiza cómo promover una estrategia europea de IA que permita hacer frente a algunos de esos retos.

a. Acción por el clima e infraestructura sostenible

123) Pese a que la lucha contra el cambio climático debería constituir una prioridad fundamental para los responsables políticos de todo el mundo, la transformación digital y la IA fiable ofrecen un potencial enorme para reducir el impacto que ejerce el ser humano sobre el medio ambiente y posibilitar un uso eficiente y eficaz de la energía y los recursos naturales[58]. La IA fiable puede, por ejemplo, combinarse con macrodatos para identificar con mayor precisión las necesidades energéticas, mejorando de ese modo la eficiencia de la infraestructura energética y del consumo de energía[59].

124) Centrándonos en sectores como el transporte público, cabe la posibilidad de utilizar sistemas de IA en los sistemas de transporte inteligentes[60] para minimizar las colas, optimizar las rutas, permitir que las personas con problemas de visión sean más independientes[61], optimizar la eficiencia energética de los motores y, de ese modo, potenciar los esfuerzos de descarbonización y reducir la huella ecológica en favor de una sociedad más respetuosa con el medio ambiente. En la actualidad, cada 23 segundos fallece una persona en el mundo como

57 https://sustainabledevelopment.un.org/?menu=1300.

58 Existen varios proyectos de la UE que buscan desarrollar las redes inteligentes y el almacenamiento de energía, que pueden contribuir al éxito de la transición energética respaldada por la tecnología, incluso a través de soluciones basadas en IA y otras soluciones digitales. Con el fin de complementar el trabajo de esos proyectos, la Comisión ha puesto en marcha la iniciativa BRIDGE, que permite a los proyectos de redes inteligentes y almacenamiento de energía Horizonte 2020 crear una visión común sobre una serie de cuestiones transversales: https://www.h2020-bridge.eu/.

59 Véase, por ejemplo, el proyecto Encompass: http://www.encompass-project.eu/.

60 Las nuevas soluciones basadas en la IA ayudan a preparar a las ciudades para la movilidad del futuro. Véase, por ejemplo, un proyecto financiado por la UE titulado Fabulos: https://fabulos.eu/.

61 Véase, por ejemplo, el proyecto PRO4VIP, que forma parte de la estrategia europea Visión 2020 dirigida a combatir la ceguera evitable, especialmente la asociada a la vejez. Una de las áreas prioritarias del proyecto era la movilidad y la orientación.

consecuencia de un accidente de tráfico[62]. Los sistemas de IA pueden ayudar a reducir considerablemente el número de personas fallecidas en accidentes, por ejemplo mediante la mejora de los tiempos de reacción y del respeto de las normas[63].

b. Salud y bienestar

125) Existe la posibilidad de utilizar tecnologías de IA fiables —y, de hecho, ya se están utilizando— para mejorar la eficiencia y la personalización de los tratamientos y para ayudar a prevenir enfermedades que ponen en peligro la vida humana[64]. Gracias a la IA, los médicos y los profesionales de la medicina pueden llevar a cabo análisis más precisos y detallados de los complejos datos sanitarios de sus pacientes incluso antes de que estos caigan enfermos, y proporcionarles un tratamiento preventivo personalizado[65]. En el contexto del envejecimiento de la población europea, la IA y la robótica pueden ser herramientas muy valiosas para ayudar a los cuidadores y contribuir al cuidado de las personas mayores[66], así como para vigilar las condiciones de los pacientes en tiempo real, lo que permitirá salvar muchas vidas[67].

126) La IA fiable también puede resultar de gran ayuda en un nivel más general. Por ejemplo, puede examinar e identificar tendencias generales en el sector de la atención y el tratamiento sanitarios[68], permitiendo así detectar enfermedades con mayor rapidez, desarrollar medicamentos de forma más eficiente, ofrecer tratamientos más personalizados[69] y, en última instancia, salvar un mayor número de vidas.

---

62 https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

63 El proyecto europeo UP-Drive, por ejemplo, pretende abordar los desafíos descritos en el ámbito del transporte realizando contribuciones que posibiliten una automatización progresiva de los vehículos y la colaboración entre ellos, facilitando así un sistema de transporte más inclusivo y más asequible. https://up-drive.eu/.

64 Véase, por ejemplo, el proyecto REVOLVER (Repeated Evolution of Cancer): https://www.healtheuropa.eu/personalised-cancer-treatment/87958/, o el proyecto Murab, que realiza biopsias más precisas y tiene el objetivo de diagnosticar el cáncer y otras enfermedades con mayor rapidez: https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

65 Véase, por ejemplo, el proyecto Live INCITE: www.karolinska.se/en/live-incite. Este consorcio de proveedores de servicios sanitarios desafía a la industria a desarrollar soluciones inteligentes basadas en la IA y en las TIC que permitan desarrollar intervenciones orientadas a mejorar la forma de vida en el proceso perioperatorio. Se pretende desarrollar soluciones innovadoras de sanidad electrónica que puedan influir en los pacientes de forma personalizada para adoptar las medidas necesarias en su estilo de vida tanto antes como después de someterse a una operación quirúrgica con objeto de optimizar el resultado desde el punto de vista sanitario.

66 El proyecto CARESSES, financiado por la UE, está relacionado con la utilización de robots para el cuidado de personas mayores, centrándose en la sensibilidad cultural de estas: los robots adaptan su forma de actuar y de hablar a la cultura y los hábitos de las personas mayores a las que asisten: http://caressesrobot.org/en/project/. Véase también la aplicación de IA llamada Alfred, un asistente virtual que ayuda a las personas mayores a mantenerse activas: https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. Además, el proyecto EMPATTICS (EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems) estudiará y definirá de qué modo pueden los profesionales sanitarios y sus pacientes utilizar las tecnologías de la información y las comunicaciones (TIC), incluidos los sistemas de IA, para planificar intervenciones con pacientes y llevar a cabo un seguimiento del progreso de su estado físico y mental: www.empattics.eu.

67 Véase, por ejemplo, el proyecto MyHealthAvatar (www.myhealthavatar.eu), que ofrece una representación digital del estado de salud de un paciente. Este proyecto de investigación desarrolló una aplicación y plataforma en línea que permite recopilar y acceder a la información digital sobre el estado de salud del paciente a largo plazo. Adopta la forma de un compañero sanitario a lo largo de toda la vida («avatar»). MyHealthAvatar también predice el riesgo de sufrir diabetes, enfermedades cardiovasculares, accidentes cerebrovasculares e hipertensión.

68 Véase, por ejemplo, el proyecto ENRICHME (www.enrichme.eu), que aborda la disminución progresiva de la capacidad cognitiva de la población a medida que envejece. Una plataforma integrada para la vida cotidiana asistida por el entorno y un robot de servicios de movilidad capaz de efectuar una labor de seguimiento a largo plazo e interactuar con las personas mayores ayudará a estas a seguir siendo independientes y permanecer activas durante más tiempo.

69 Véase, por ejemplo, la utilización de la IA por parte de Sophia Genetics, basado en la explotación de la inferencia estadística, el reconocimiento de patrones y el aprendizaje automático para maximizar el valor de los datos de la genómica y la radiómica: https://www.sophiagenetics.com/home.html.

c. Educación de calidad y transformación digital

127) Los cambios tecnológicos, económicos y medioambientales obligan a la sociedad a ser más proactiva. Gobiernos, líderes industriales, instituciones educativas y sindicatos se enfrentan a la responsabilidad de ayudar a los ciudadanos a realizar la transición a la nueva era digital, garantizando que posean las cualificaciones que requerirán los puestos de trabajo del futuro. Las tecnologías de la IA fiable podrían ayudar a predecir con mayor exactitud qué puestos de trabajo y qué profesiones se verán afectados de un modo fundamental por la tecnología, qué nuevas funciones se crearán y qué competencias se necesitarán. Esto podría ayudar a los gobiernos, sindicatos y a la industria a planificar la (re)cualificación de los trabajadores. También podría ofrecer una vía para el reciclaje profesional a aquellos ciudadanos que teman que sus cualificaciones puedan quedar obsoletas.

128) Además, la IA puede ser una herramienta magnífica para combatir las desigualdades en el ámbito de la enseñanza y crear programas educativos personalizados y adaptables que podrían ayudar a que cualquier persona adquiera nuevas cualificaciones, competencias y aptitudes en función de su propia capacidad de aprendizaje[70]. Esto podría aumentar tanto la velocidad del aprendizaje como la calidad de la educación, desde la escuela primaria hasta la universidad.

2. Ejemplos de preocupaciones fundamentales que plantea la IA

129) Cuando se incumple uno de los componentes de la IA fiable, surge una preocupación fundamental en relación con la IA. Muchas de las preocupaciones que se enumeran a continuación estarán ya recogidas en el ámbito de aplicación de los requisitos legales existentes, que son de obligado cumplimiento. Sin embargo, incluso en circunstancias en las que se haya demostrado el cumplimiento de los requisitos legales, es posible que estos no abarquen todas las preocupaciones que pueden surgir desde el punto de vista ético. Dado que nuestra comprensión de la pertinencia de las normas y los principios éticos evoluciona sin cesar y puede cambiar a lo largo del tiempo, es posible que la lista siguiente (que no es exhaustiva) se abrevie, amplíe, modifique o actualice en el futuro.

a. Identificación y seguimiento de personas mediante la IA

130) La IA posibilita una identificación cada vez más eficiente de las personas físicas por parte tanto de entidades públicas como privadas. Entre los ejemplos destacables de una tecnología escalable de identificación mediante la IA cabe citar el reconocimiento facial y otros métodos involuntarios de identificación a través del uso de datos biométricos (detectores de mentiras, evaluación de la personalidad utilizando microexpresiones, detección automática de la voz...). En ocasiones, la identificación de las personas es el resultado deseable y está en consonancia con los principios éticos (por ejemplo en el caso de la detección del fraude, el blanqueo de capitales o la financiación del terrorismo). No obstante, la identificación automática plantea serias preocupaciones tanto desde el punto de vista legal como ético, dado que puede tener efectos inesperados en muchos niveles psicológicos y socioculturales. Para preservar la autonomía de los ciudadanos europeos es necesario, por tanto, utilizar las técnicas de control de la IA de manera proporcionada. Una definición clara de si (y, en su caso, cuándo y de qué manera) se puede utilizar la IA con fines de identificación automática de personas, diferenciando entre la identificación de una persona frente a su seguimiento y rastreo, y entre una vigilancia selectiva o masiva, será crucial para hacer realidad una IA fiable. La aplicación de este tipo de tecnologías debe estar claramente justificada en la legislación existente[71]. Cuando la base jurídica para llevar a

[70] Véase, por ejemplo, el proyecto MaTHiSiS, cuyo objetivo es proporcionar una solución para el aprendizaje basado en el afecto en un entorno confortable; dicha solución se apoya en complejos algoritmos y dispositivos tecnológicos (http://mathisis-project.eu/). Véase también el proyecto «Watson va a clase» de IBM o la plataforma de Century Tech.

[71] En este sentido, cabe recordar el artículo 6 del RGPD, que establece, entre otras cosas, que el tratamiento de datos únicamente será lícito si cuenta con una base jurídica válida.

cabo dicha actividad sea el «consentimiento», deberán desarrollarse medios prácticos[72] que permitan identificar automáticamente a través de tecnologías de IA o equivalentes el consentimiento real y verificado que se deba otorgar. Esto también es aplicable al uso de datos personales «anónimos» que sea posible asociar posteriormente a personas concretas.

b. Sistemas de IA encubiertos

131) Los seres humanos siempre deberían saber si están interactuando directamente con otro ser humano o con una máquina. Los responsables de garantizarlo son los profesionales de la IA. Dichos profesionales deberían, por tanto asegurar que las personas sean conocedoras de que están interactuando con un sistema de IA o puedan preguntar y dar su aprobación al respecto (por ejemplo, mediante la publicación de cláusulas de exención de responsabilidad claras y transparentes). Obsérvese que existen casos limítrofes que complican la cuestión (por ejemplo, una voz emitida por un ser humano pero filtrada a través de un sistema de IA). Es preciso tener presente que la confusión entre humanos y máquinas puede tener múltiples consecuencias, como el establecimiento de vínculos, la influencia o la reducción del valor de la persona[73]. El desarrollo de robots humanoides[74] debería ser objeto, por tanto, de una evaluación pormenorizada desde el punto de vista ético.

c. Evaluación de ciudadanos mediante IA vulnerando los derechos fundamentales

132) Las sociedades deben esforzarse por proteger la libertad y la autonomía de todos los ciudadanos. Cualquier forma de evaluación de los ciudadanos puede dar lugar a una pérdida de autonomía de estos y poner en peligro el principio de no discriminación. Este tipo de evaluación solamente debe utilizarse si existe una justificación clara; en todo caso, las medidas empleadas deben ser proporcionadas y justas. La evaluación normativa de los ciudadanos (un análisis general de su «personalidad moral» o de su «integridad ética») en *todos* los aspectos y a gran escala por parte de las autoridades públicas o de agentes privados pone en peligro esos valores, especialmente cuando no se aplica de conformidad con los derechos fundamentales o cuando se utiliza de manera desproporcionada y sin un propósito legítimo claramente definido y comunicado.

133) En la actualidad, la evaluación de los ciudadanos —ya sea a gran o a pequeña escala— suele usarse en evaluaciones puramente descriptivas y en ámbitos muy concretos (por ejemplo, en los sistemas escolares, en el campo del aprendizaje electrónico o para la expedición de permisos de conducir). Incluso en este tipo de aplicaciones tan específicas se debería poner a disposición de los ciudadanos un proceso plenamente transparente que incluya información sobre el proceso, la finalidad y la metodología de la evaluación. Téngase en cuenta que la transparencia no puede impedir la discriminación ni garantizar la equidad; tampoco es la panacea contra el problema de la evaluación. Lo ideal sería ofrecer a los ciudadanos la posibilidad de dejar de estar sometidos al mecanismo de evaluación cuando sea posible sin sufrir consecuencia alguna; de lo contrario, se deberán ofrecer mecanismos para impugnar y rectificar las evaluaciones. Esto reviste una importancia especial en situaciones en las que exista asimetría de poder entre las partes. Se debería garantizar la disponibilidad de este tipo de opciones de salida en el diseño de la tecnología en circunstancias en las que sea necesario para garantizar el cumplimiento de los derechos fundamentales, algo que es necesario en una sociedad democrática.

d. Sistemas de armas letales autónomas

134) En la actualidad, un número desconocido de países e industrias están investigando y desarrollando sistemas de armas letales autónomas, desde misiles capaces de fijar blancos selectivos hasta máquinas con aptitudes

72 Tal como muestran los mecanismos actualmente utilizados para otorgar el consentimiento informado en internet, los consumidores suelen dar su consentimiento sin un análisis detallado. Por lo tanto, estos mecanismos difícilmente pueden calificarse de prácticos.

73 Madary y Metzinger (2016). *Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology.* Frontiers in Robotics and AI, 3(3).

74 Esto también es aplicable a los avatares que funcionan mediante inteligencia artificial.

cognitivas y de aprendizaje, con el fin de decidir cuándo, dónde y contra quién combatir sin intervención humana. Esto plantea problemas fundamentales desde el punto de vista ético, ya que puede dar lugar a una carrera armamentística incontrolable hasta un nivel sin precedentes en la historia y crear contextos militares en los que prácticamente no exista control humano y no se aborden los riesgos de un funcionamiento inadecuado. El Parlamento Europeo ha instado a desarrollar con urgencia una posición común y jurídicamente vinculante para tratar las cuestiones éticas y legales relacionadas con el control humano, la supervisión, la rendición de cuentas y la aplicación de la legislación internacional de derechos humanos, el Derecho internacional humanitario y las estrategias militares[75]. Recordando el objetivo de la Unión Europea de fomentar la paz, consagrado en el artículo 3 del Tratado de la Unión Europea, defendemos y apoyamos la resolución del Parlamento de 12 de septiembre de 2018 y todas las iniciativas relacionadas con los sistemas de armas letales autónomas.

e. Preocupaciones que pueden surgir a largo plazo

135) El desarrollo de la IA todavía está vinculado a campos concretos y requiere de científicos e ingenieros humanos adecuadamente formados que especifiquen sus objetivos con precisión. No obstante, si nos proyectamos hacia el futuro con un horizonte temporal más amplio, cabe imaginar algunas preocupaciones cruciales (aunque hipotéticas) a largo plazo[76]. Un enfoque basado en el riesgo sugiere la necesidad de tenerlas en cuenta, en vista de los posibles «desconocidos desconocidos» y «cisnes negros»[77]. La naturaleza de estas preocupaciones, cuyo impacto puede ser muy importante, unida a la incertidumbre actual sobre el rumbo de los acontecimientos, exige evaluar periódicamente estas cuestiones.

D. CONCLUSIÓN

136) En este documento se recogen las directrices éticas sobre la IA elaboradas por el Grupo de expertos de alto nivel sobre inteligencia artificial.

137) Reconocemos el efecto positivo que ya ejerce y seguirá ejerciendo la IA, tanto desde el punto de vista comercial como social. No obstante, también nos preocupa garantizar que los riesgos y otros efectos adversos asociados a estas tecnologías se gestionen de manera adecuada y proporcionada con arreglo a la aplicación de la IA. La IA es una tecnología tanto transformadora como disruptiva, y su evolución a lo largo de los últimos años se ha visto facilitada por la disponibilidad de enormes cantidades de datos digitales, grandes avances tecnológicos en el ámbito de la capacidad informática y de almacenamiento y significativas innovaciones científicas y en el campo de la ingeniería en relación con los métodos y herramientas de la IA. Los sistemas de IA continuarán afectando a la sociedad y a los ciudadanos de formas que todavía no podemos ni siquiera imaginar.

138) En este contexto, es importante desarrollar sistemas de IA merecedores de confianza, puesto que los seres humanos solamente podrán confiar en ellos y aprovechar todos los beneficios que ofrece si tanto esta tecnología como las personas y los procesos subyacentes a ella son fiables. Por lo tanto, la elaboración de estas directrices responde a la aspiración de crear una IA fiable.

139) La IA fiable tiene tres componentes: 1) debe ser lícita y cumplir todas las leyes y reglamentos aplicables; 2) ha de ser ética, de modo que se garantice el respeto de los principios y valores éticos, y 3) debe ser robusta tanto desde el punto de vista técnico como social, a fin de asegurar que los sistemas de IA, incluso si las intenciones son buenas, no provoquen daños accidentales. Cada uno de estos componentes es necesario pero no

---

75 Resolución 2018/2752(RSP) del Parlamento Europeo.

76 Pese a que hay quien considera que la inteligencia artificial general, la conciencia artificial, los agentes morales artificiales, la superinteligencia o la IA transformadora pueden ser ejemplos de este tipo de preocupaciones a largo plazo (que no existen en la actualidad), otras muchas personas piensan que estos ejemplos son poco realistas.

77 Un evento de tipo «cisne negro» es un suceso muy poco frecuente, aunque con elevado impacto; tan raro, que puede que ni siquiera sea observado. Por lo tanto, la probabilidad de que ocurra normalmente solo puede estimarse con un alto grado de incertidumbre.

suficiente para el logro de una IA fiable. Lo ideal es que todos ellos actúen en armonía y de manera simultánea. Cuando surjan tensiones, deberíamos esforzarnos por resolverlas.

140) En el capítulo I se articulan los derechos fundamentales y un conjunto de principios éticos asociados que resultan cruciales en el contexto de la IA. En el capítulo II enumeramos siete requisitos clave que deberían cumplir los sistemas de IA para hacer realidad la IA fiable. En dicho capítulo se proponen métodos técnicos y no técnicos que pueden contribuir a su aplicación. Por último, en el capítulo III se incluye una lista de evaluación de la IA fiable que puede ayudar a poner en práctica esos siete requisitos. En la última sección del documento se ofrecen ejemplos de las oportunidades beneficiosas y preocupaciones cruciales que plantean los sistemas de IA, sobre los que esperamos estimular un debate en mayor profundidad.

141) Europa goza de una ventaja única basada en un enfoque consistente en situar al ciudadano en el centro de todos sus esfuerzos. Este enfoque está integrado en el propio ADN de la Unión Europea, a través de los Tratados en los que se sustenta. El documento actual forma parte de una visión que promueve la IA fiable, que entendemos debería ser la base sobre la que Europa desarrolle una posición de liderazgo en el terreno de la inteligencia artificial mediante la creación de sistemas de IA innovadores y de vanguardia. Esta ambiciosa visión ayudará a garantizar la prosperidad de los ciudadanos europeos, tanto individual como colectiva. Nuestro objetivo es crear una cultura de «IA fiable para Europa», en la que los beneficios de la IA lleguen a todos los ciudadanos de un modo que garantice el respeto de nuestros valores fundacionales: los derechos fundamentales, la democracia y el Estado de Derecho.

GLOSARIO

142) El presente glosario forma parte de las directrices; su propósito es ayudar a comprender los términos que se utilizan en este documento.

Sistemas de inteligencia artificial o IA

143) Los sistemas de inteligencia artificial (IA) son sistemas de software (y en algunos casos también de hardware) diseñados por seres humanos[78] que, dado un objetivo complejo, actúan en la dimensión física o digital mediante la percepción de su entorno a través de la obtención de datos, la interpretación de los datos estructurados o no estructurados que recopilan, el razonamiento sobre el conocimiento o el procesamiento de la información derivados de esos datos, y decidiendo la acción o acciones óptimas que deben llevar a cabo para lograr el objetivo establecido. Los sistemas de IA pueden utilizar normas simbólicas o aprender un modelo numérico; también pueden adaptar su conducta mediante el análisis del modo en que el entorno se ve afectado por sus acciones anteriores.

144) La IA es una disciplina científica que incluye varios enfoques y técnicas, como el aprendizaje automático (del que el aprendizaje profundo y el aprendizaje por refuerzo constituyen algunos ejemplos), el razonamiento automático (que incluye la planificación, programación, representación y razonamiento de conocimientos, búsqueda y optimización) y la robótica (que incluye el control, la percepción, sensores y accionadores así como la integración de todas las demás técnicas en sistemas ciberfísicos).

145) El Grupo de expertos de alto nivel sobre inteligencia artificial ha elaborado un documento separado en el que profundiza en la definición de *sistemas de IA* utilizada a efectos de este documento. Dicho documento se ha publicado en paralelo con el título «Una definición de la inteligencia artificial: Principales capacidades y disciplinas científicas».

Profesionales de la IA

146) Entendemos por profesionales de la IA todas aquellas personas u organizaciones dedicadas al desarrollo (incluidas las labores de investigación, diseño o suministro de datos para), despliegue (incluida la aplicación) o utilización de sistemas de IA, salvo aquellas que utilicen sistemas de IA en calidad de usuarios finales o consumidores.

Ciclo de vida de los sistemas de IA

147) El ciclo de vida de un sistema de IA abarca las fases de desarrollo (incluidas las tareas de investigación, diseño, provisión de datos y realización de ensayos limitados), despliegue (incluida la aplicación) y utilización de dicho sistema.

Auditabilidad

148) La auditabilidad se refiere a la capacidad de un sistema de IA de someterse a la evaluación de sus algoritmos, datos y procesos de diseño. Constituye uno de los siete requisitos que debería cumplir cualquier sistema de IA fiable. Esto no implica necesariamente que siempre deba disponerse de forma inmediata de la información sobre los modelos de negocio y la propiedad intelectual del sistema de IA. El hecho de garantizar la existencia de mecanismos de trazabilidad y registro desde las primeras fases de diseño del sistema de IA puede favorecer la auditabilidad del sistema.

Sesgo

149) Un sesgo es una inclinación que favorece o perjudica a una persona, objeto o posición. En los sistemas de IA pueden surgir numerosos tipos de sesgos. Por ejemplo, en los sistemas de IA impulsados por datos, como los creados a través del aprendizaje automático, los sesgos en la recogida de datos y la formación pueden dar

[78] Los seres humanos diseñan sistemas de IA directamente, aunque también pueden emplear técnicas de IA para optimizar su diseño.

lugar a sesgos en el sistema de IA. En los sistemas de IA lógicos, como los basados en normas, pueden surgir sesgos como consecuencia de la visión que puede tener un ingeniero del conocimiento acerca de las reglas aplicables en un entorno específico. También pueden aparecer sesgos debido a la formación y adaptación en línea a través de la interacción, o como consecuencia de la personalización en aquellos casos en que se presentan a los usuarios recomendaciones o información adaptadas a sus gustos. Los sesgos no tienen por qué estar relacionados necesariamente con inclinaciones humanas o con la recogida de datos por parte de personas. Pueden surgir, por ejemplo, en los limitados contextos en los que se utiliza un sistema, en cuyo caso no existe la posibilidad de generalizarlo a otros contextos. Los sesgos pueden ser positivos o negativos, intencionados o no. En algunos casos, pueden dar lugar a resultados discriminatorios o injustos, lo que en este documento se denomina «sesgo injusto».

Ética

150) La ética es una disciplina académica que constituye un subcampo de la filosofía. En general, se ocupa de cuestiones como «¿qué es una buena acción?», «¿qué valor tiene la vida humana?», «¿qué es la justicia?» o «¿qué es una buena vida?». En el ámbito de la ética académica, existen cuatro campos principales de investigación: i) la metaética, que se ocupa fundamentalmente del significado y la referencia de los enunciados normativos, y cómo se pueden determinar sus valores de verdad (si es que los tienen); ii) la ética normativa, que es el modo práctico de determinar un curso moral de acción mediante el examen de las normas relativas a los actos correctos o incorrectos y la asignación de un valor a determinadas acciones; iii) la ética descriptiva, que busca investigar desde el punto de vista empírico el comportamiento y las creencias morales de las personas; y iv) la ética aplicada, a la que le preocupa lo que estamos obligados a hacer (o lo que se nos permite hacer) en una situación específica (a menudo nueva desde el punto de vista histórico) o en un ámbito determinado de posibilidades de acción (a menudo sin precedentes históricos). La ética aplicada se ocupa de situaciones de la vida real, en las que es necesario tomar decisiones bajo limitaciones de tiempo y, a menudo, de racionalidad. La ética de la inteligencia artificial se considera generalmente un ejemplo de la ética aplicada que se centra en los problemas normativos que plantea el desarrollo, despliegue y utilización de la IA.

151) En los debates éticos se utilizan con frecuencia los términos «moral» y «ético». El primero se refiere a lo concreto, las pautas de comportamiento, las costumbres y convenciones que se pueden encontrar en determinadas culturas, grupos o personas en un momento específico. El término «ético», por su parte, hace referencia a una evaluación de esas acciones y comportamientos concretos desde una perspectiva sistemática y académica.

IA ética

152) En este documento se utiliza la expresión «IA ética» para indicar el desarrollo, despliegue y utilización de la IA de un modo que garantice el cumplimiento de las normas éticas, incluidos los derechos fundamentales como derechos morales especiales, los principios éticos y los valores esenciales asociados. Es el segundo de los tres elementos clave necesarios para hacer realidad una IA fiable.

IA centrada en la persona

153) Una IA con un enfoque centrado en la persona se esfuerza por asegurar que los valores humanos ocupen un lugar central en el desarrollo, despliegue, utilización y supervisión de los sistemas de IA, garantizando el respeto de los derechos fundamentales, incluidos los recogidos en los Tratados de la Unión Europea y en la Carta de los Derechos Fundamentales de la Unión Europea; todos ellos constituyen una referencia unitaria a un fundamento común arraigado en el respeto de la dignidad humana, en el que el ser humano disfruta de una condición moral única e inalienable. Esto requiere asimismo tener en cuenta el entorno natural y el resto de seres vivos que forman parte del ecosistema humano, así como un enfoque sostenible que permita la prosperidad de las generaciones futuras.

Equipos rojos

154) Se denomina «equipos rojos» a la práctica en la que un grupo independiente desafía a una organización a mejorar su eficacia, asumiendo para ello un papel o punto de vista contradictorio. Este método se utiliza, en particular, para ayudar a identificar y hacer frente a posibles vulnerabilidades que afecten a la seguridad.

Reproducibilidad

155) La reproducibilidad describe si un experimento con IA muestra el mismo comportamiento cuando se repite varias veces en las mismas condiciones.

IA robusta

156) La solidez de un sistema de IA abarca tanto su solidez técnica (que resulta adecuada en un contexto determinado, como el ámbito de aplicación o la fase del ciclo de vida) así como desde el punto de vista social (garantizando que el sistema de IA tenga debidamente en cuenta el contexto y el entorno en el que opera). Esto es crucial para asegurar que, incluso si las intenciones son buenas, el sistema no provoque daños involuntarios. La solidez es el último de los tres componentes necesarios para hacer realidad una IA fiable.

Partes interesadas

157) Entendemos por partes interesadas todas aquellas dedicadas a la investigación, desarrollo, diseño, despliegue o utilización de la IA, así como aquellas que se ven afectadas de forma directa o indirecta por esta, incluidas, con carácter no limitativo, empresas, organizaciones, investigadores, servicios públicos, instituciones, organizaciones de la sociedad civil, gobiernos, autoridades reguladoras, interlocutores sociales, personas físicas, ciudadanos, trabajadores y consumidores.

Trazabilidad

158) La trazabilidad de un sistema de IA se refiere a su capacidad para llevar a cabo un seguimiento de los datos, el desarrollo y el proceso de despliegue del sistema, generalmente a través de un proceso de identificación y registro documentados.

Confianza

159) En este documento se adopta la presente definición tomada de la bibliografía: «La confianza se considera como: 1) un conjunto de creencias específicas que tienen que ver con la benevolencia, la competencia, la integridad y la previsibilidad (creencias confiadas); 2) la voluntad de una parte de depender de otra en una situación de riesgo (intención confiada); o 3) la combinación de los elementos anteriores»[79]. Pese a que la confianza quizá no sea una propiedad atribuible a las máquinas, en este documento se hace hincapié en la importancia de poder confiar no solo en el hecho de que los sistemas de IA cumplan las leyes y los principios éticos y sean sólidos, sino también en poder confiar en todas las personas y procesos involucrados en el ciclo de vida de los sistemas de IA.

Una IA fiable

160) La IA fiable tiene tres componentes: 1) debe ser lícita, es decir, cumplir todas las leyes y reglamentos aplicables; 2) ha de ser ética, demostrando el respeto y garantizando el cumplimiento de los principios y valores éticos, y 3) debe ser robusta, tanto desde el punto de vista técnico como social, puesto que los sistemas de IA, incluso si las intenciones son buenas, pueden provocar daños accidentales. La fiabilidad de la IA no concierne únicamente a la fiabilidad del propio sistema de inteligencia artificial, sino también a la de todos los procesos y agentes implicados en el ciclo de vida del sistema.

Personas y grupos vulnerables

79 Siau, K., Wang, W. (2018), «Building Trust in Artificial Intelligence, Machine Learning, and Robotics», *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), S. 47–53.

161) Debido a su heterogeneidad, no existe una definición generalmente aceptada ni que cuente con un consenso amplio del concepto de «personas vulnerables». Lo que se considera una persona o grupo vulnerable suele depender del contexto. Los sucesos vitales de carácter temporal (como la infancia o la enfermedad), los factores de mercado (como la asimetría de información o el poder de mercado), los factores económicos (como la pobreza), los vinculados a la identidad de las personas (como el género, la religión o la cultura) y otros pueden desempeñar un papel en ese sentido. La Carta de los Derechos Fundamentales de la Unión Europea recoge en su artículo 21, relativo a la no discriminación, los motivos de discriminación siguientes, que pueden servir como punto de referencia, entre otros: el sexo, la raza, el color, los orígenes étnicos o sociales, las características genéticas, la lengua, la religión o las convicciones, las opiniones políticas o de cualquier otro tipo, la pertenencia a una minoría nacional, el patrimonio, el nacimiento, la discapacidad, la edad o la orientación sexual. En las disposiciones de otras leyes se abordan los derechos de determinados grupos, además de los enumerados anteriormente. Este tipo de listas nunca pueden ser exhaustivas, y pueden cambiar a lo largo del tiempo. Un grupo vulnerable es un grupo de personas que comparten una o varias características de vulnerabilidad.

Este documento ha sido elaborado por los miembros del Grupo de expertos de alto nivel sobre inteligencia artificial,

que se relacionan a continuación en orden alfabético:

Pekka Ala-Pietilä, presidente del Grupo de expertos de alto nivel
AI Finland, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann – Coponente
Zalando
Mária Bieliková
Universidad eslovaca de Tecnología, Bratislava
Cecilia Bonefeld-Dahl – Coponente
DigitalEurope
Yann Bonnet
ANSSI

Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
Iniciativa del IEEE para la Ética de los Sistemas Inteligentes/Autónomos y Universidad de la Sorbona
Mark Coeckelbergh
Universidad de Viena
Virginia Dignum – Coponente
Universidad de Umeå
Luciano Floridi
Universidad de Oxford
Jean-Francois Gagné – Coponente
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Agencia de los Derechos Fundamentales
Sami Haddadin
Escuela de Robótica de Múnich y MI
Gry Hasselbalch
Laboratorio de ideas DataEthics y Universidad de Copenhague
Fredrik Heintz
Universidad de Linköping
Fanny Hidvegi
Access Now
Eric Hilgendorf
Universidad de Wurzburgo
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
TU Wien
Robert Kroplewski
Abogado y asesor del Gobierno de Polonia
Elisabeth Ling
RELX

Pierre Lucas
Orgalim – Industrias tecnológicas europeas

Ieva Martinkenaite
Telenor
Thomas Metzinger – Coponente
JGU Mainz y Asociación Europea de Universidades
Catelijne Muller
ALLAI Netherlands y Comité Económico y Social Europeo (CESE)
Markus Noga
SAP
Barry O'Sullivan, vicepresidente del Grupo de expertos de alto nivel
Colegio Universitario de Cork
Ursula Pachl
BEUC
Nicolas Petit – Coponente
Universidad de Lieja
Christoph Peylo
Bosch

Iris Plöger
BDI
Stefano Quintarelli
Garden Ventures
Andrea Renda
College of Europe Faculty y CEPS
Francesca Rossi
IBM
Cristina San José
Federación Bancaria Europea
George Sharkov
Digital SME Alliance
Philipp Slusallek
Centro Alemán de Investigación en IA (DFKI)
Françoise Soulié Fogelman
Consultora de IA

Saskia Steinacker – Coponente
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe – Coponente
TU Delft
Thiébaut Weber
Confederación Europea de Sindicatos (CES)
Cecile Wendling
AXA
Karen Yeung – Coponente
Universidad de Birmingham

Urs Bergmann, Cecilia Bonefeld-Dahl, Virginia Dignum, Jean-François Gagné, Thomas Metzinger, Nicolas Petit, Saskia Steinacker, Aimee Van Wynsberghe y Karen Yeung actuaron como ponentes para la elaboración de este documento.

Pekka Ala-Pietilä preside el Grupo de expertos de alto nivel sobre inteligencia artificial. Su vicepresidente, Barry O'Sullivan, se encarga de la coordinación del segundo entregable del Grupo. Nozha Boujemaa, vicepresidenta hasta el 1 de febrero de 2019, coordinó el primer entregable y realizó aportaciones al contenido de este documento.

Nathalie Smuha brindó apoyo editorial.