La Comisión Europea

GRUPO DE EXPERTOS DE ALTO NIVEL SOBRE INTELIGENCIA ARTIFICIAL

PROYECTO DE DIRECTRICES ÉTICAS SOBRE UNA IA CONFIABLE

RESUMEN

Documento de trabajo para la consulta de las partes interesadas

Bruselas, 18 de diciembre de 2018

PROYECTO DE DIRECTRICES ÉTICAS

SOBRE UNA IA CONFIABLE

Grupo de expertos de alto nivel sobre inteligencia artificial
Proyecto de Directrices éticas sobre una IA confiable

Comisión Europea
Dirección General de Comunicación

Persona de contacto: Nathalie Smuha (Coordinadora del Grupo de expertos de alto nivel sobre inteligencia artificial)
Correo electrónico: CNECT-HLG-AI@ec.europa.eu

Comisión Europea
B - 1049 Bruselas

Documento publicado el 18 de diciembre de 2018, en inglés.

El grupo de expertos de alto nivel sobre inteligencia artificial elaboró este documento de trabajo sin perjuicio de la postura individual de sus miembros sobre determinadas cuestiones y de la versión final. Se seguirá trabajando en este documento y se presentará la versión final en marzo de 2019, una vez haya concluido la consulta de las partes interesadas a través de la Alianza europea de IA.

Ni la Comisión Europea ni cualquier persona que actúe en su nombre serán responsables del uso que pudiera hacerse de esta información. El contenido de este documento de trabajo es responsabilidad exclusiva del Grupo de expertos de alto nivel sobre inteligencia artificial (AI HLEG). Aunque el personal de los servicios de la Comisión ha facilitado la preparación de las Directrices, los puntos de vista expresados reflejan la opinión del AI HLEG y, en ningún caso, pueden considerarse como una postura oficial de la Comisión Europea. Este es un borrador del primer entregable del AI HLEG cuya versión final se presentará a la Comisión en marzo de 2019. La versión final del segundo entregable (Política sobre inteligencia artificial y recomendaciones de inversión) se presentará a mediados de 2019.

Hay más información disponible acerca del Grupo de expertos de alto nivel sobre inteligencia artificial online (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

material no sujeto a los derechos de autor de la UE, debe solicitarse permiso directamente a los titulares de los derechos de autor.

RESUMEN

Este documento de trabajo constituye un borrador de las Directrices éticas sobre inteligencia artificial elaborado por el Grupo de expertos de alto nivel sobre inteligencia artificial (AI HLEG), cuya versión final está prevista para marzo de 2019.

La inteligencia artificial (IA) es una de las fuerzas más transformadoras de nuestro tiempo y está destinada a modificar el tejido social. Supone una gran oportunidad para aumentar la prosperidad y el crecimiento que Europa debe tratar de lograr. Durante la última década, se han llevado a cabo grandes avances gracias a la enorme cantidad de datos digitales y a la potente arquitectura informática disponibles, así como a los avances en técnicas de IA tales como el aprendizaje automático. Grande innovaciones en vehículos autónomos, atención sanitaria, sistemas domóticos, educación o ciberseguridad, posibles gracias a la IA, están mejorando la calidad de nuestra vida cotidiana. Además, la IA es fundamental para abordar muchos de los grandes retos a los que se enfrenta el mundo, como la salud y el bienestar mundial, el cambio climático, sistemas democráticos y jurídicos fiables, y otros expresados en los objetivos de desarrollo sostenible de las Naciones Unidas.

Aunque la IA es capaz de generar enormes beneficios para las personas y la sociedad, también entraña riesgos que se deben gestionar de manera adecuada. Puesto que, en general, los beneficios de la IA compensan sus riesgos, debemos seguir un que **maximice los beneficios y minimice los riesgos**. Para asegurarnos de que vamos por el buen camino, es necesario regirse por un **enfoque de la IA centrado en los seres humanos**. Es decir, que nos obligue a recordar que el desarrollo y uso de la IA tienen por objetivo mejorar el bienestar de los seres humanos, y no verlos como un medio en sí mismos. **La IA confiable marcará nuestro camino**, ya que los seres humanos solo podrán beneficiarse completamente y con plena confianza de la IA si pueden confiar en la tecnología.

La IA confiable tiene **dos componentes**: 1) debe respetar los derechos fundamentales, las leyes vigentes y los principios y valores esenciales, de manera que se garantice un «**fin ético**», y 2) debe ser fiable y **sólida técnicamente** hablando, ya que un escaso dominio tecnológico puede provocar daños involuntarios, aunque las intenciones sean buenas.

Por tanto, estas Directrices establecen el **marco de una IA confiable**:

- El c**apítulo I** tiene por objeto **garantizar el fin ético de la IA**, ya que establece los derechos fundamentales, así como los principios y valores esenciales, que debe cumplir.
- Teniendo en cuenta esos principios, en el c**apítulo II** se formulan las **directrices para la consecución** de una IA confiable, abordando tanto el fin ético como la solidez técnica. Se enumeran los requisitos de una IA confiable, y se proporciona un resumen de los métodos técnicos y no técnicos que se pueden usar para su aplicación.
- En el c**apítulo III** se aborda la **aplicación** de los requisitos, ya que se proporciona una lista concreta, aunque no exhaustiva, de aspectos que se deben evaluar para conseguir una IA confiable. A continuación, esta lista se adapta a casos prácticos específicos.

Al contrario que otros documentos que versan sobre la IA ética, estas Directrices no pretenden facilitar otra lista de principios y valores esenciales, sino servir de guía en cuanto a su aplicación y funcionamiento real en sistemas de IA. Se ofrece orientación en tres niveles de abstracción: desde el más abstracto en el capítulo I (derechos, principios y valores fundamentales) al más concreto en el capítulo III (lista de aspectos para evaluar).

Estas Directrices van dirigidas a todas las **partes implicadas que desarrollan, aplican o usan la IA**, abarcando a empresas, organizaciones, investigadores, servicios públicos, instituciones, individuos u otras entidades. En la versión final se expondrá un mecanismo que permita a las partes interesadas

secundar las directrices.

Cabe señalar que, con estas Directrices, no se pretende sustituir ninguna elaboración de políticas ni regulación (asunto tratado en el segundo entregable de la AI HLEG: Política y recomendaciones de inversión, previsto para mayo de 2019), ni tampoco impedir su introducción sino que deben considerarse como un documento en constante evolución que necesita actualizaciones periódicas, ya que la tecnología y nuestro conocimiento sobre ella evolucionan constantemente. Por tanto, este documento es un punto de partida para el debate sobre la «**IA confiable hecha en Europa**».

Aunque Europa solo puede difundir su enfoque ético en materia de IA cuando es competitiva a escala mundial, este **es fundamental para que la competencia sea responsable**, ya que generará confianza en el usuario y facilitará una acogida de la IA más amplia. Estas Directrices no están pensadas para reprimir la innovación en materia de IA en Europa, sino para usar la ética como inspiración en el desarrollo de una marca propia de IA, una marca que proteja y ayude a las personas y al bien común. De este modo, Europa podrá posicionarse como líder en una IA de vanguardia, segura y ética. Los ciudadanos europeos solo aprovecharán los beneficios de la IA si se garantiza la confianza.

Por último, más allá de Europa, otro objetivo de estas Directrices es **incentivar la reflexión y el debate** sobre el marco ético en materia de IA a nivel mundial.

ORIENTACIÓN EJECUTIVA

En cada capítulo de las Directrices, se ofrece orientación para la consecución de una IA confiable dirigida a todas las partes interesadas que desarrollen, apliquen o usen la IA. A continuación se expone un resumen:

Capítulo I: Orientaciones clave para garantizar el fin ético:

- Garantizar que la IA se centre en los seres humanos: La IA debe desarrollarse, aplicarse y usarse con un **fin ético**, basado en los derechos fundamentales, los valores sociales y los principios éticos de *beneficencia* (hacer el bien), *no maleficencia* (no hacer daño), *autonomía de los seres humanos*, *justicia* y *explicabilidad*. Esto es fundamental a la hora de trabajar para conseguir una **IA confiable**.
- Ampararse en los derechos fundamentales, los principios y valores éticos para evaluar de forma prospectiva los posibles efectos de la IA en los seres humanos y el bien común. Prestar **especial atención** a situaciones que afectan a grupos vulnerables como niños, personas con discapacidad o minorías, o a situaciones con **asimetrías de poder o información**, como las que hay entre empleadores y empleados, o empresas y consumidores.
- Ser consciente de que la IA, aunque aporte importantes ventajas para las personas y la sociedad, puede tener un impacto negativo. Permanecer alerta ante cuestiones de especial preocupación.

Capítulo II: Orientaciones clave para lograr una IA confiable:

- Incorporar los **requisitos para lograr una IA confiable** desde la primera fase de diseño: rendición de cuentas, gestión de datos, diseño universal, gestión de la autonomía de la IA (supervisión humana), no discriminación, respeto de la autonomía de los seres humanos, respeto de la privacidad, robustez, seguridad y transparencia.
- Plantear métodos técnicos y no técnicos que garanticen la aplicación de esos requisitos en el sistema de IA. Además, tener en cuenta esos requisitos a la hora de constituir el equipo que se ocupe del sistema, crear el propio sistema, establecer el entorno de prueba y trabajar en sus posibles aplicaciones.
- Proporcionar **información a las partes interesadas** (clientes, empleados, etc.), de una manera clara

y proactiva, sobre las posibilidades y limitaciones del sistema de IA, de modo que permita establecer expectativas realistas. A este respecto es fundamental garantizar la **trazabilidad** del sistema de IA.

- Conseguir que la IA confiable forme parte de la cultura de las organizaciones, y proporcionar información a las partes interesadas sobre cómo aplicarla en el diseño y el uso de los sistemas de IA. También se puede incluir la IA confiable en los capítulos deontológicos o los códigos de conducta de las organizaciones.
- Garantizar la participación y la **inclusión de las partes interesadas** en el diseño y el desarrollo de los sistemas de IA. Además, procurar que haya **diversidad** a la hora de establecer los equipos que van a desarrollar, aplicar y probar el producto.
- Esforzarse por facilitar la auditabilidad de los sistemas de IA, concretamente en situaciones o contextos críticos. En la medida de lo posible, diseñar el sistema de modo que se pueda hacer un seguimiento de las decisiones individuales sobre las diferentes contribuciones, datos, modelos previamente constituidos, etc. Definir también los **métodos de explicación** del sistema de IA.
- Garantizar un proceso específico para la **gestión de la rendición de cuentas**.
- Prever acciones de **educación y formación**, así como garantizar que los gestores, desarrolladores, usuarios y empleadores conozcan la IA confiable y estén formados al respecto.
- Tener en cuenta que puede haber enormes conflictos entre los diferentes objetivos (la transparencia puede abrir la puerta a un uso indebido; identificar y corregir los errores puede oponerse a la protección de la privacidad). Transmitir información al respecto y documentarse.
- Promover la investigación y la innovación de modo que favorezca la consecución de los requisitos para lograr una IA confiable.

Capítulo III: Orientaciones clave para evaluar la IA confiable

- Adoptar una lista de aspectos que se deben evaluar a la hora de desarrollar, aplicar o usar la IA para que esta sea fiable, así como adaptar esta lista a los casos prácticos concretos en los que se use la IA.
- Tener en cuenta que este tipo de listas **nunca serán exhaustivas**, y que tratar de conseguir una IA confiable no es marcar casillas, sino un proceso continuo en el que identificar requisitos, evaluar soluciones y mejorar los resultados en todo el ciclo de vida del sistema de IA.

Estas orientaciones forman parte de un enfoque de la IA que se centra en el ser humano y que le permitirá a Europa convertirse en líder mundial en innovación en materia de IA de vanguardia, ética y segura. Su objetivo es facilitar una **«IA confiable hecha en Europa»** que contribuirá al bienestar de los ciudadanos europeos.