Trustworthy AI

Trustworthy AI Card Deck

Co-funded by the
Erasmus+ Programme
of the European Union

Este proyecto ha sido financiado con el apoyo de la Comisión Europea. Esta publicación [comunicación] refleja únicamente las opiniones del autor, y la Comisión no se hace responsable del uso que pueda hacerse de la información contenida en ella.

Sección 1

Introducción

La IA Fiable

La IA tiene el potencial de fortalecer la sociedad y apoyar a los seres humanos en diferentes frentes. Sin embargo, debemos implementar cuidadosamente los sistemas de IA para proteger nuestros derechos fundamentales, nuestra salud y nuestra seguridad. Desarrollar, desplegar y utilizar la IA de forma centrada en el ser humano puede aportarnos una IA segura, protegida, justa e inclusiva. Para lograrlo, el Grupo de Expertos de Alto Nivel en Inteligencia Artificial elaboró las Directrices Éticas para una IA Fiable[1] . Como base para estas Directrices, el grupo eligió tres pilares que deben sustentar la IA para que sea Fiable:

1. debe ser legal, lo que debe cumplir con todas las leyes y reglamentos aplicables
2. debe adherirse a los principios y valores éticos, y
3. debe ser sólida desde el punto de vista técnico y social y no causar daños involuntarios

Para los pilares 2 y 3, el grupo elaboró **7 requisitos para una IA Fiable** que debería guiar el desarrollo, el despliegue y el uso de la IA en Europa:

1. Agencia Humana y Supervisión: Los sistemas de IA deben apoyar la autonomía humana y permitirles tomar decisiones informadas. Para lograrlo, los sistemas de IA deben actuar como facilitadores de una sociedad democrática y equitativa, apoyando la agencia del usuario, fomentando los derechos fundamentales y permitiendo la supervisión humana. Hay que enseñar a los estudiantes los niveles adecuados de agencia y autonomía humanas, el control humano, y la dignidad humana en general.

2. Robustez técnica y seguridad: Este principio exige que los sistemas de IA se desarrollen con un enfoque preventivo de los riesgos y de manera que se comporten de forma fiable según lo previsto, minimizando los daños involuntarios e inesperados y evitando los daños inaceptables. Además, debe garantizarse la integridad física y mental de los seres humanos. Los estudiantes deben aprender a reconocer y garantizar la precisión y fiabilidad de los sistemas de IA. Además, los estudiantes deben saber cómo equilibrar la solidez técnica y las limitaciones éticas.

3. Privacidad y gobernanza de los datos: La privacidad es un derecho fundamental especialmente afectado por los sistemas de IA. La prevención del daño a la privacidad debe ser una prioridad. Este requisito requiere, lógicamente, una gobernanza de datos adecuada. Esto abarca la calidad e integridad de los datos utilizados, su relevancia considerando el ámbito en el que se desplegarán los sistemas de IA, sus protocolos de acceso y la capacidad de procesar los datos de manera que se proteja la privacidad. Hay que enseñar a los estudiantes a recopilar y reconocer datos de alta calidad, a manejarlos con sensibilidad, a mantener la privacidad y a evitar sesgos en los datos y en los modelos construidos a partir de ellos.

4. Transparencia: Este requisito tiene dos partes. Abarca la transparencia de los elementos relevantes para los sistemas de IA; los datos recogidos, la formación y el funcionamiento del sistema, explicaciones de resultados y los modelos comerciales pertinentes. También abarca la obligación de ser transparente sobre el uso de los sistemas de IA y no utilizarlos de forma encubierta. Los estudiantes deben reconocer sistemas transparentes, y adquirir habilidades para desarrollar una IA explicable. Esto implica enseñar a los estudiantes a documentar y comunicar adecuadamente el uso de los datos, así como las decisiones tomadas en el proceso de diseño.

[1] Directrices éticas para una IA digna de confianza, Grupo de Expertos de Alto Nivel en Inteligencia Artificial, 2019

5. Diversidad, no discriminación y equidad: Los resultados de los sistemas de IA deben ser no discriminatorios y estar libres de prejuicios inaceptables. Otra parte importante de este requisito es la igualdad de acceso mediante procesos de diseño inclusivos. La inclusión y la diversidad deben estar presentes durante todo el ciclo de vida del sistema de IA. Esto incluye la consideración y la participación de todas las partes interesadas afectadas a lo largo del proceso. Los estudiantes deben aprender sobre la importancia y el valor añadido de la experiencia interdisciplinaria a la hora de desarrollar, desplegar y utilizar los sistemas de IA. También se les debe enseñar los posibles efectos discriminatorios de las decisiones tomadas a lo largo del proceso de desarrollo.

6. Bienestar medioambiental y social: El medio ambiente y la sociedad en su conjunto deben considerarse "partes interesadas" durante el ciclo de vida del sistema. Este requisito incluye el fomento de la sostenibilidad y la responsabilidad ecológica. Implica tanto la investigación de soluciones de IA que aborden el cambio climático u otras preocupaciones sociales, como ser conscientes de la huella ecológica de la formación y el despliegue de un sistema de IA. También implica comprender y mitigar los efectos sociales, democráticos o sistémicos más amplios que puede tener la IA.

7. Responsabilidad: Este requisito requiere que se establezcan mecanismos para garantizar la responsabilidad y la rendición de cuentas de los sistemas de IA y sus resultados, tanto antes como después de su desarrollo, despliegue y uso. Se debe enseñar a los estudiantes sobre la auditoría y el mantenimiento de registros, así como los marcos legales para la responsabilidad y poder demostrar la minimización de los efectos negativos.

Estos requisitos deben abordarse y evaluarse continuamente a lo largo del ciclo de vida del sistema de IA -desde la fase de diseño y desarrollo hasta el final de su uso-, teniendo en cuenta métodos tanto técnicos como no técnicos para garantizar su cumplimiento. De este modo, se puede fomentar y garantizar una IA ética y sólida.

¿Por qué enseñar sobre la IA Fiable?

Es probable que muchos estudiantes participen en el desarrollo, la implantación, la adquisición o el uso de la IA en sus futuros trabajos. Por lo tanto, es importante que adquieran las habilidades necesarias para desarrollar e implementar la IA de manera responsable y fiable. Por ello, hemos transformado los 7 requisitos para una IA Fiable en ejercicios específicos y recursos educativos que ayudan a enseñar habilidades específicas, entre ellas la capacidad de:

1. Identificar la aplicabilidad de los 7 requisitos en diferentes contextos y sus diferentes dimensiones para las diferentes partes interesadas
2. Deliberar sobre las posibles aplicaciones de los 7 requisitos
3. Seleccionar y aplicar un curso de acción en respuesta a un análisis ético sobre los requisitos

Al adquirir estas habilidades y desarrollar la capacidad de comprender y actuar de acuerdo con los valores éticos y sociales, los estudiantes pueden garantizar un enfoque "humano al mando" de la IA, en el que sigue dependiendo de ellos decidir si, cuándo y cómo deben desarrollarse, desplegarse y utilizarse los sistemas de IA. Dado que la responsabilidad de una IA Fiable no recae sólo en los desarrolladores, sino en diferentes partes interesadas y en expertos de otros campos, la enseñanza a los estudiantes de diferentes ámbitos es de gran importancia.

IA Fiable para Estudiantes de STEM

Es fácil imaginar que los alumnos que estudian asignaturas STEM (matemáticas, biología, física, química, etc.) pueden acabar formando parte del desarrollo de la IA. A menudo surgen debates en torno a las consecuencias no deseadas de la IA, el sesgo algorítmico, la recogida y protección de datos, etc. Muchos de estos problemas surgen durante el desarrollo. Por eso es importante que los estudiantes de STEM, los "futuros desarrolladores de IA", sepan apreciar los valores éticos y que los apliquen en las diferentes etapas del desarrollo. Esto les permitirá identificar en una fase temprana las consecuencias imprevistas o relaciones causales entre determinadas opciones de desarrollo y sus problemas éticos. Por ejemplo, pueden asegurarse de que la recogida de datos se haga de forma adecuada y responsable, minimizando el riesgo de resultados sesgados.

Todo desarrollador/a de IA o investigador/a de IA estudiará una o más materias STEM en algún momento de su vida. Por lo tanto, es importante que la ética esté bien integrada en su carrera educativa. No sólo para prevenir las consecuencias imprevistas del desarrollo, sino para adquirir conocimientos fiables que permitan tomar decisiones informadas.

IA Fiable para Estudiantes de Administración de Empresas

La enseñanza de la IA de confianza en el ámbito de la Administración de Empresas es importante dado el crecimiento exponencial de la IA en entornos empresariales, su uso comercial, y el rápido crecimiento de las empresas tecnológicas. Muchos consumidores están expuestos a la IA a través de chatbots de atención al cliente, sistemas de predicción/recomendaciones que toman decisiones sobre ellos, o dispositivos IoT o IoB (Internet of Bodies) que utilizan la IA para su propio funcionamiento. Con mayor frecuencia, el reclutamiento, la contratación y la evaluación de los trabajadores dentro de las organizaciones se realizan con IA. Además, las empresas pueden beneficiarse de las del análisis de datos que la IA puede ofrecer. La IA puede ser muy beneficiosa para la logística, por ejemplo, en ayudar a determinar el mejor sistema organizativo para los vuelos. También puede ayudar en el sector sanitario revisando los historiales médicos, los enfoques de los tratamientos y/o complementar los conocimientos de los médicos. El aumento del uso de la IA en la vida cotidiana de las personas también hace que sea atractiva para añadir la IA a la línea de productos de las empresas.

Aunque es evidente que la IA ofrece muchas oportunidades, las empresas deben tener cuidado con los riesgos que llevan falta de seguridad, los sesgos e injusticias, pero también el desplazamiento masivo de puestos de trabajo o el impacto negativamente en la sociedad o la democracia. No hay que pasar por alto estos riesgos sólo por las posibilidades económicas que puede ofrecer la IA.

IA de confianza para Estudiantes de Ciencias Políticas

Como hemos comentado hasta ahora, no basta con que la IA se desarrolle de forma fiable. También es crucial que se despliegue de forma responsable. Por desgracia, no podemos depositar toda nuestra confianza en las empresas y los desarrolladores para que esto ocurra. Las políticas y los marcos legislativos establecen los límites dentro de los cuales la IA puede desarrollarse y utilizarse de forma aceptable. Por eso es importante que los estudiantes de ciencias políticas reciban también formación sobre la IA fiable. Si asumimos que muchos de estos estudiantes se convertirán en responsables políticos o tendrán un papel en la gobernanza, es esencial que entiendan las verdaderas capacidades y limitaciones de la IA y cómo la política y la legislación pueden establecer los límites adecuados para la IA.

Sección 2

La Baraja de Cartas

Cartas de la AI Fiable

El objetivo de esta baraja de cartas es crear un debate ético significativo en clase y comprender la diversidad de aplicaciones de la IA en diferentes ámbitos. Con esta baraja, los alumnos comprenderán la complejidad de algunos dilemas éticos y su relación con los puntos de vista de las diferentes partes interesadas. Las tarjetas ofrecen una explicación general de las técnicas de IA que pueden mezclarse y combinarse en diferentes ámbitos para crear un caso de uso. La falta de familiaridad con las técnicas de IA puede provocar ciertas dificultades con algunos ejercicios, por lo que también hemos creado una baraja de tarjetas de casos de uso preseleccionados para facilitarlo. También somos conscientes de que no todas las técnicas de IA están presentes en esta baraja. Las técnicas y los enfoques de la IA (y sus nombres) evolucionan constantemente y se están desarrollando conforme hablamos. Por lo tanto, por ahora, simplemente hemos presentado un conjunto de técnicas de IA fáciles de entender para que puedan utilizar por personas de diferentes ámbitos.

Antes de utilizar la baraja de cartas, aconsejamos que los alumnos vean los clips de conocimiento para familiarizarse con el concepto de IA Fiable y sus 7 requisitos. No obstante, cualquier conocimiento adicional sobre la IA será útil para plantear los casos de uso necesarios para algunos de los ejercicios. Puede obtener una visión de las técnicas generales de IA en el siguiente enlace: ¿Qué es la Inteligencia Artificial? En 5 minutos (en inglés)

Composición de la baraja

La baraja consta de 5 conjuntos de cartas:

Con estas cartas los estudiantes pueden inspirarse de casos de uso para pensar en cómo utilizar diferentes técnicas de IA en diferentes ámbitos. Utilizando las cartas de requisitos, pueden analizar críticamente las técnicas y los casos de uso desde el punto de vista de diferentes partes interesadas. Todo esto puede hacerse a través de una serie de ejercicios de clase que se explica a continuación.

- **Técnica de IA:** Una técnica de IA genérica que puede utilizarse en muchos ámbitosde diversas maneras
- **Ámbito:** Un sector en el que se puede aplicar una técnica de IA que contiene múltiples subsectores pertenecientes al mismo.
- **Requisito:** Un requisito ético que evalúa la fiabilidad de una técnica de IA aplicada en un ámbito específico. Estas tarjetas también pueden utilizarse para evaluar las cartas de casos de uso.
- **Parte interesada:** Papel de una persona en el desarrollo o despliegue de un sistema de IA que tiene su propio "interés competitivo", por ejemplo, el dinero, la eficiencia, la seguridad, la equidad, la privacidad, la autonomía, etc.
- **Caso de uso:** Un ejemplo del funcionamiento y los objetivos de una técnica de IA aplicada a un **Técnica de IA**
- intervienen varias partes interesadas.

10 Técnicas de IA, con ejemplos

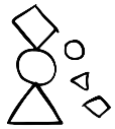
8 Requisitos (7 Requisitos de la IA Fiable + Derechos Fundamentales)

14 Ámbitos, con distintas áreas

7 Partes Interesadas, con ejemplos

13 Casos de Uso

Ejercicios discusión



Ejercicio 1 - Defienda su caso de uso

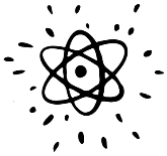
Se necesitan: Cartas de ámbito, cartas de técnica, cartas de partes interesadas, cartas de requisitos
6 jugadores: 1 juez, 5 desarolladores

1. Se roba una carta del mazo de ámbito. Este será el ámbito en esa ronda.
2. Cada jugador roba y muestra una carta de interesado.
3. Cada jugador roba una carta de técnica sin mostrarla.
4. Cada jugador diseña un caso de uso de la IA (sobre el papel) dentro del ámbito desde el punto de vista de su parte interesada.
5. **NB: para una versión más sencilla: que cada jugador coja una tarjeta de caso de uso y diseña el sistema de IA descrito en ese caso de uso**
6. Una vez que todos los jugadores hayan diseñado su caso de uso, el juez sacará una tarjeta de requisitos por estudiante.
7. El juez pregunta al primer desarrollador si su caso de uso cumple con el requisito y, en caso de que si, cómo.
8. Si la respuesta es aceptable, el desarrollador recibe 2 puntos.
9. Si no es aceptable, o si el estudiante no puede responder, cualquiera de los otros desarrolladores puede argumentar por qué su caso cumple con el requisito.
10. Si esta respuesta es aceptable, el otro desarrollador recibe 2 puntos.
11. El juez repite los pasos 7-10 para cada uno de los promotores.
12. El diseño con más puntos es discutido por todos los desarrolladores para evaluar los resultados.

Ejercicio 2 - Diseño y debate en grupo

Se necesitan: Cartas de técnica, cartas de ámbito, cartas de partes interesadas, cartas de requisitos
5 jugadores

1. En grupo, elijan una tarjeta de ámbito y otra de técnica.
2. Juntos, inventen un caso de uso para ese ámbito y esa técnica.
3. Cada jugador coge una carta de parte interesada.
4. Tome una tarjeta de requisito a la vez y comience a discutir ese requisito desde la perspectiva de su parte interesada.
5. Repita el paso 4 para cada requisito.
6. Opcional: Ordene las tarjetas de requisitos por orden de prioridad desde el punto de vista de sus propias partes interesadas y comente por qué ha elegido esa orden.



Ejercicio 3 - Discutir un caso de uso

Se necesitan: Cartas de caso de uso, cartas de partes interesadas, cartas de requisitos

5 jugadores

1. En pequeños grupos, elija una tarjeta de caso de uso.
2. Cada jugador coge una carta de parte interesada.
3. Tome una tarjeta de requisito a la vez (5 en total) y comience a discutir ese requisito desde la perspectiva de su parte interesada.
4. Añade: Ordene las tarjetas de requisitos por orden de prioridad desde el punto de vista de sus propias partes interesadas y comente por qué ha elegido esa orden.

Las Partes Interesadas

Una parte interesada es el papel de una persona en el desarrollo o la implantación de un sistema de IA que tiene sus propios "intereses contrapuestos", por ejemplo, el dinero, la eficiencia, la seguridad, la equidad, la privacidad, la autonomía, *etc*. Entendemos que la representación de algunas de estas partes interesadas puede resultar difícil para algunos, por lo que ofrecemos algunos ejemplos para que sirvan de inspiración.

Gobernanza: Los diferentes responsables gubernamentales pretenden garantizar que la IA sea transparente, sostenible y cumpla con los diferentes requisitos éticos y legales. Interés: cumplimiento legal

Autoridad/supervisor: Las autoridades supervisoras son autoridades independientes que velan por la aplicación y el cumplimiento uniforme y coherente de las normas y leyes (dentro de su ámbito específico de competencia). Interés: seguridad

Implantador: El implantador de un sistema de IA tiene como objetivo garantizar que el sistema funcione de forma rentable y cumpla con sus objetivos. Interésese: eficiencia, dinero

Afectado: Los afectados se preocupan por su propio bienestar (físico y mental) y por la protección de sus derechos humanos fundamentales. Intereses: equidad, privacidad, autonomía

Experto en el ámbito: los expertos en el campo tienen como objetivo informar a los desarrolladores, a los implantadores y a cualquier otra parte interesada con información correcta y fiable sobre su campo de experiencia para garantizar que un sistema funcione correctamente tanto a nivel técnico como social. Intereses: equidad, seguridad, información correcta

Desarrollador: El objetivo de los desarrolladores es desarrollar sistemas que funcionen correctamente teniendo en cuenta los objetivos del implantador. Interés: eficiencia, seguridad

Las Cartas

Las siguientes paginas contienen la baraja de cartas completa en dos versiones:

1. Versión plegable (para imprimir de una cara)
2. Versión a dos caras (para imprimir a dos caras)